The London School of Economics and Political Science

Philosophical Issues in Evidence-Based Medicine: Evaluating the Epistemological Role of Double Blinding and Placebo Controls

Jeremy Howick

A thesis submitted to the Department of Philosophy, Logic, and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, March 2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I certify that the thesis I have presented for examination for the PhD degree to the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Abstract

The Evidence-Based Medicine (EBM) movement endorses a hierarchy of evidence that places randomized controlled trials at the top. More specifically, double-blind, placebo-controlled trials are often considered to be the 'best of the best'. This view leads to the paradox that treatments that seem to be most strongly supported by evidence, ranging from tracheotomies to rabies vaccines, have never been tested in randomized trials of any description and are hence supported by (allegedly) sub-optimal evidence. Moreover many of these treatments do not seem supportable by best evidence – how, for example do we keep the surgeons who perform tracheotomies 'blind'? After a brief introduction (chapter 1), and review of the literature (chapter 2), I argue that criticisms of the EBM hierarchy can be launched from the simple basis that best evidence rules out the most plausible rival hypothesis (chapter 3). To examine the relative evidential weight of placebo controlled trials compared to 'active' controlled trials (in which the control treatment is an existing established treatment) requires a good deal of conceptual work. I defend a modified version of Grünbaum's (1981/1986) definition of placebos (chapter 4), then provide constraints on what can count as a 'legitimate' placebo control (chapter 5). Next, I explain why double-blinding does not always rule out additional rival hypotheses. I then argue that the arguments for the superiority of placebo controls are flawed. The 'assay sensitivity' argument is limited in scope and based on a misconception about the nature of placebo controls (chapter 7), while the claim that only placebo controlled trials measure the absolute effect size relies on the questionable assumption that placebo and non-placebo effects add rather than interact (chapter 8). I conclude that the evidence hierarchy endorsed by EBM does not stand on solid foundations.

Acknowledgments

This thesis is dedicated to Kor, Jack, and Raven.

Many people helped me write this thesis. Many of the ideas for the thesis were developed during conversations with my primary supervisor, John Worrall. I credit John with most of the good ideas. John was also especially helpful in the final stages of writing up. My other supervisor, Nancy Cartwright, provided invaluable insights and advice at many key stages. She was also very helpful in San Diego where I spent a term studying under her at the UCSD. On a few occasions, Nancy stopped me from going insane and encouraged me to persevere. Roman Frigg, who also played a supervisory role in the early stages of my thesis, gave invaluable advice about the more technical aspects of my thesis. Discussions with several fellows at the Centre for the Philosophy of the Natural and Social Sciences, including Damien Fennell and Phillip Thonemann, were also helpful. Sharing ideas with Rupert Sheldrake was enlightening. Several research students were also involved in developing my ideas. These include Sheldon Steed, Lefteris Farmakis, Gary Jones, Foad Dizadji-Bahmani and Kizito Kiyimba. Writing the thesis also required support from family and friends. The unconditional love of my mother kept me on track. My father and younger sisters Samantha, Katie, and Teresa will be happy and proud that I've finished. Many long-time friends, including Mark, Sebastien, John, and Sam, where always there when I needed them. Frances was generous with her spare room. Mingy, Stephen, Carolyn and Dr. Bali all played a vital role. My examiners, Professor Donald Gillies and Professor Richard Ashcroft gave me extremely helpful comments; they were also very encouraging. Thanks to anyone I forgot to mention.

Table of Contents

Table of Figures

Chapter One. Overlooked Problems with the Received View that Placebo Controlled, Double-Blind, Randomized Trials Provide the 'Best Evidence'

> *[M]ost scientists and EBM advocates are ignorant of the philosophy of science and give little or no thought to constructing a philosophical basis for their activities. ... One hopes that the attention of philosophers will be drawn to these questions*
>
> *- (Haynes 2002)*

The Evidence-Based Medicine (EBM) movement, which enjoys widespread support, endorses a hierarchy of evidence that places the randomized, controlled trial (RCT) at the top[1]. More specifically, double blind, placebo controlled randomized trials are often considered to be the 'best of the best'. Very briefly (much more will follow), a randomized trial is a medical experiment where participants are divided into experimental and control groups by some random process analogous to flipping a coin. Then, the participants who are allocated to the experimental group get the experimental treatment. Then, those who are allocated to the control group receive the control treatment, which is usually either an existing established treatment, or a 'placebo' (for example, a sugar pill delivered in the belief that it could be the experimental treatment). A double blind, or double masked trial is one in which neither the participants in the trial nor those delivering the intervention are aware of who receives the experimental intervention and who receives the control.

The problem with the EBM view is that many of the treatments in whose effectiveness we have the most confidence – that we consider to be most strongly supported by evidence, have never been tested in randomized trials of any description. These treatments include Automatic External Defibrillation to start a stopped heart, tracheostomy to open a blocked air passage, the Heimlich manoeuvre to dislodge an obstruction in the breathing passage, rabies vaccines, penicillin for the treatment of

---

[1] The view that RCTs provide the best evidence is most strongly propounded by the Evidence-Based Medicine (EBM) movement (Straus, Richardson, and Haynes 2005), but also enjoys widespread support by the regulatory bodies around the world (Harbour 2008; Phillips et al. 2001; Canadian Task Force on the Periodic Health Examination 1979; US Preventive Services Task Force 1996). Sometimes, the claim is that the 'gold standard' is a collection of systematically analyzed RCTs ('systematic reviews'). For simplicity, I will ignore systematic reviews for the present chapter, and address the issue in some detail in the next.

pneumonia, and adrenaline injections to treat severe anaphylactic shock. If the current view is correct, then our confidence in these seemingly effective treatments is exaggerated, perhaps irrational. Meanwhile, we often lack confidence in some treatments that are supported by 'best' evidence (i.e. double blind, placebo controlled RCTs). The antidepressant 'Prozac', for instance, has proven superior to placebo in some double blind RCTs, yet the effects of Prozac (over and above 'placebo' effects) have been disputed (Kirsch and Sapirstein 1998; Kirsch and Moore 2002; Healy 2004; Healy 2006; Moncrieff and Kirsch 2005). Exploiting this irony, Gordon Smith and Jill Pell, wrote a spoof article entitled "Parachute use to prevent death and major trauma related to gravitational challenge: a systematic review of randomised controlled trials". They concluded that:

> Advocates of evidence-based medicine have criticised the adoption of interventions evaluated by using only observational [i.e. not from RCTs] data. We think that everyone might benefit if the most radical protagonists of evidence based medicine organized and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute" (Smith and Pell 2003).

If RCTs provide the 'best' evidence, then how do we account for the greater confidence sometimes placed in treatments that have never been tested in RCTs?

More relevantly for the purposes of this thesis, it seems indeed very difficult to see how, practically and ethically speaking, treatments like tracheostomies *could* be subjected to placebo controlled, double-blind RCTs and hence garner 'best evidence' in their favour. In fact, many of the treatments that have never been tested in any RCTs, simply could not be tested in double blind, placebo controlled conditions. What would count as a placebo control for a tracheostomy? How could we possibly keep the surgeons blind? The view that double blind, placebo controlled RCTs provide the best evidence leads to an *a priori* judgement that treatments which are untestable in double blind and placebo controlled conditions are unsupportable by 'best' evidence.

Ironically, treatments that cannot be tested in double blind conditions include those that are most effective: if a treatment is dramatically effective, both participants and investigators will quickly and correctly guess that it is not a placebo. For example, if a placebo controlled trial of a new drug to treat severe anaphylactic shock were ever approved by ethics committees, and the new injection were even more effective than the current treatments (such as epinephrine), neither the patient nor the physician would suspect even for a minute that it was a placebo. The paradox that the most effective

treatments cannot be supported by 'best' (double blind) evidence is known as the 'Phillip's Paradox' (Ney, Collins, and Spensor 1986).

This raises the issue of whether the judgments about weight of evidence embodied in the evidence hierarchy are sound. As will become clear, much attention (from both philosophers and non-philosophers alike), has been paid to the issue of whether RCTs in general provide the best evidence[2] and whether randomization in particular is an epistemic good[3]. Yet the relative evidential weight of double blinding and placebo controls have escaped careful scrutiny. Of course, the debate over whether double blind, placebo controlled RCTs are superior to other RCTs is intertwined with the wider question of whether RCTs in general provide the best evidence. Nobody claims that a small and very badly conducted RCT provides better evidence than other studies. Rather, only 'high quality' RCTs are said to provide the best evidence. 'High quality', is taken to mean, among other things, that the trial was performed in 'double blind'[4] conditions, and sometimes, that it was placebo controlled[5]. The view that double blinding increases the methodological quality of a study has gone unquestioned, while in spite of two notable exceptions (Kirsch 2000; Anderson 2006), the methodological superiority of placebo controlled trials has been taken for granted.

In this thesis I will propose a more fundamental standard of evidence, namely the view that best evidence rules out more rival hypotheses than other evidence. Of course the theory of weight of evidence, aka confirmation theory, has always been

---

[2] In the philosophical literature, see for example (Worrall 2007b, 2007a, 2002; Cartwright 2007; Grossman and MacKenzie 2005; Borgerson 2005; Bluhm 2005; Ashcroft 2004). In medical literature see for example (Cochrane 1972; Barton 2000; Pocock and Elbourne 2000; Benson and Hartz 2000; Penston 2003; Brighton et al. 2003; Higgins 2005).

[3] Supporters of randomization include (Papineau) and (Pearl), while critics include many Bayesian philosophers of science (Urbach 1985; Lindley 1993, 1982; Howson and Urbach 1993), and more recently, Worrall (2002, 2007a).

[4] Claims that double blinding adds epistemic value can be found in virtually every medical textbook (Straus, Richardson, and Haynes 2005, p.122; Armitage, Berry, and Matthews 2002, p. 605; Bland 2000, p. 19; Hill and Hill 1991, p. 214; Greenhalgh 2006, p. 66; Jadad 1998, p. 20).

[5] See, for example (Kaptchuk; Temple and Ellenberg; ICH; Gomberg-Maitland, Frison, and Halperin 2003; Hwang and Morikawa)

central to the philosophy of science; there has been a great deal of work on it and much remains to be settled. However I argue that criticisms of the current view of evidence can be launched from a very simple basis – from a claim that is bound to be common to any detailed account of weight of evidence. This is the claim that best evidence for some theoretical claim T, at the same time as being accounted for by T, also rules out the most plausible rival hypotheses to T. I show in chapter 3 that the accounts of evidence provided by J.S. Mill (1843), Popper (1969), and Bayesian philosophers, although superficially distinct, converge on this fundamental point.

1.1. What is to come

My first task will be to provide a more detailed account of the view that double blind, placebo controlled RCTs provide the 'best' evidence (chapter 2). I will then defend the view that a critique of the EBM position can be launched from the simple idea that best evidence rules out the most plausible rivals (chapter 3). I then turn to the issue of whether placebo controlled trials are in any way to be preferred in terms of the weight of evidence they provide to so-called active controlled trials (in which the control treatment is that currently judged most effective). This requires, I argue, a good deal of work to clarify the notions of placebos (chapter 4) and placebo controls (chapter 5). In spite of a flurry of literature[6], an adequate conceptualization of the placebo has not been provided; without such definitions, the relative advantages of placebo over active controls cannot be analyzed. I then explain why in many cases double-blinding

---

[6] Conceptual work on the placebo, although inconclusive, has been given much attention (Shapiro and Shapiro 1997; Shapiro and Morris 1978; Greenwood 1997; Grünbaum 1986, 1981; Waring 2003; Feinstein 2002, 1980; de Craen, Tijssen, and Kleijnen 1997; de Craen et al. 2000; de Craen et al. 1996). Likewise, the ethical debate over the use of placebo controls has been given much attention (WMA 2001, 1964; Weijer and Miller 2004; Weijer and Glass 2002; Freedman, Weijer, and Glass 1996; Freedman, Glass, and Weijer 1996; Emanuel and Miller 2001; Miller and Brody 2002; Ackerman 2002; Lewis et al. 2002) Discussions of the mechanism of action of the placebo have also been written about (Kirsch 2004; Moerman and Jonas 2002; Moerman 2000, 1983; Montgomery and Kirsch 1997; Evans 2003; Amanzio et al. 2001; Benedetti et al. 2003; Benedetti et al. 2004; Morris 1997). Similarly, estimates of the magnitude of the placebo effect are relatively common (Rohsenow and Marlatt 1981; Kienle and Kiene 1997; Huskisson 1974; Beecher 1962, 1961, 1955; Benson and McCallie 1979; Hróbjartsson and Gøtzsche 2004b, 2004a, 2001; Hróbjartsson 1996).

does not rule always out additional rival hypotheses (chapter 6). I then argue that the two main arguments for the superiority of placebo controls over active controls are flawed. I contend that the 'assay sensitivity argument' is both limited in scope based on a misconception about the nature of placebo controls (chapter 7). I then consider the claim that only placebo controlled trials provide a measure of absolute effect size and argue that it depends on the questionable assumption that placebo and non-placebo effects add rather than interact (chapter 8). I conclude (chapter 9) that the current view, or indeed any hard-and-fast rules of evidence are, at best, rules of thumb that must be governed by the basic idea that good evidence rules out rival hypotheses.

2. Chapter Two. The Double Blind, Placebo Controlled Trial to the Rescue: Attempts to Overcome Problems in Determining if a Treatment 'Caused' the Cure

> *Let us examine the placebo somewhat more critically... since it and 'double blind' have reached the status of fetishes in our thinking and literature*
>
> - *(Lasagna 1955)*

To learn whether vitamin C cures the common cold, we might administer vitamin C to someone who has a cold and see if their cold disappears. But even if the patient did recover, this would count as very little evidence for the effectiveness of vitamin C – for one thing the patient might, obviously, have recovered spontaneously. The ideal way to rule out this possibility would be to have access to what would happen in the counterfactual situation where the same person at the same time was withheld vitamin C; but this is obviously impossible.

As a surrogate for the impossible, people who take vitamin C – the 'experimental' or 'test' group, can be compared with people who do not – the 'control' group. The problem with this solution is that there are, at least in principle, innumerable differences between any two people or groups that could affect recovery from the common cold. Average general level of health, the average virulence of their colds, or whether they take more exercise, could all differ from the other group. There are, at least in principle, an infinite number of differences between two groups that could conceivably account for recovery from a cold that have nothing whatsoever to do with vitamin C. Collectively, differences between groups at the outset of a study are known as 'baseline differences' and a trial that has baseline differences is said to suffer from 'selection bias'[7].

---

[7] The term 'selection bias' is used in several ways. Worrall interprets the term *selection bias* as the role of the investigators in allocating participants to experimental or control groups:

> If the clinicians running a trial are allowed to determine the arm to which a particular patient is assigned then, whenever they have views about the comparative merits and comparative risks of the two treatments, there is some leeway for those clinicians to affect the outcome of the trial (Worrall 2007b`, p. 121).

At least one influential textbook follows Worrall's usage:

> *selection bias*, may arise since a knowledge or suspicion of the treatment to be used for the next participant may affect the investigator's decision whether or not to admit the participant to [the treatment group] of a trial (Armitage, Berry, and Matthews 2002, p. 600).

Although it is of course possible to state in advance some of the baseline differences that may be relevant, it is *impossible* to know for sure whether or how such differences influence the outcome. Selection bias seems to present, at least *prima facie*, an almost insurmountable barrier to determining whether the outcome of a study is due to the experimental intervention (i.e. vitamin C) or some baseline difference between the groups. In short, each baseline difference presents a rival hypothesis for the outcome: the difference between experimental and control groups could be the experimental treatment *or* a baseline difference (i.e. average level of 'health').

Even if we could, somewhat miraculously, make the control and experimental groups identical in all relevant respects 'at the baseline' further problems must be overcome before safely concluding that a different outcome in the two groups is due to vitamin C. If a person believes that he is being given the latest and 'best' treatment, then quite independently of whether the treatment is effective, the belief *could* produce positive outcomes. Contrariwise, if someone believes he is not getting the test treatment then his negative expectations could have negative effects. These different expectations could make the test treatment appear effective – even if, were these expectations equal in the two groups, the test treatment would not appear so effective. Thus if we gave vitamin C to one group and withheld it from a second, identical group, and the first group's cold disappeared on average more quickly, we *still* could not be sure that vitamin C caused the speedier recovery. The real cause of the differential recovery rates could well have been the dissimilar expectations.

Moreover, if the physicians know which people receive the test treatment, they might, for example, lavish more attention on them. Or, on the contrary, believing that those taking the test treatment do not need special attention, might spend more of their

---

This usage, however, is not universal. 'Selection bias' can also refer to the way participants are accepted or rejected for a trial (Jadad 1998, p. 30), or to the way entire studies are accepted or rejected for review (Green and Higgins 2005). Most commonly, however, selection bias refers to *any* systematic baseline difference between comparison groups, whether they are due to the attitudes of the investigators or not (Green and Higgins 2005). I will adopt the arbitrary convention of employing selection bias in the last way, and will replace what Worrall and some others refer to as *selection bias* (i.e. bias introduced either consciously or unconsciously by the allocators) with *allocation bias*.

scarce time with the other patients. I will call bias introduced by the investigators in charge of dispensing the intervention (the 'dispensers') that arises from knowledge of who receives the experimental intervention *dispenser bias.*

The differing treatment, like differing expectations of the participants, could make it difficult to distinguish between the effects of participant expectations and dispenser bias and the 'true' effects of the test treatment when the 'playing field is level'. Differences between the people in the test and control groups that arise once the study begins (i.e. once they have been given or withheld a test treatment such as vitamin C) are often referred to as 'performance bias'[8]. Differing participant expectations and dispenser bias are the main sources of performance bias[9].

Together, selection bias (resulting from baseline differences) and performance bias (that arises after the treatment is administered) present major obstacles to be overcome before determining whether an intervention, such as vitamin C, is effective for the cure of some ailment, such as the common cold. To use the language of the trade, there are too many potential 'confounders', or 'confounding factors'. A confounder is a source of any bias, including selection and performance bias[10]. More specifically, a confounder is a factor that has the following three properties:

---

[8] Like 'selection bias', the term 'performance bias' is not universal so my choice is partly arbitrary. Although the term 'performance bias' is used in the way I describe by some authors (Greenhalgh 2006`, p.67), other authors use the term 'ascertainment bias' to describe bias introduced by knowledge of which intervention each participant is receiving (Jadad 1998`, p. 32; Bland 2000`, p. 38).

[9] Another common source of performance bias is drop outs. If, for instance, those participants in the experimental group who suffer from severe side effects withdraw from the trial, the trial could be biased. The comparison at the end of the trial would be between those who 'survived' the experimental treatment with the (in this instance more complete) control group. Such a comparison is obviously unfair. Although it is always possible to have more withdrawals in one group rather than the other, the chances of a biased drop out rate increases when patients know the difference between treatment and control groups. Differential drop out rates are allegedly solved by an 'intention-to-treat' analysis, whereby all patients are analyzed in the groups to which they were allocated, whether or not they actually completed the course of treatment. I will not evaluate the 'intention-to-treat' strategy here.

[10] In addition to selection and performance bias, bias can arise when the outcomes are assessed, analyzed, and disseminated.

1. The factor is unrelated to the experimental intervention.
2. The factor is a determinant of the outcome.
3. The factor is unequally distributed between experimental and control groups (Straus, Richardson, and Haynes 2005`, p. 181).

It is often claimed that randomized controlled trials, or RCTs do not suffer from confounding of any kind. Briefly (more is to follow below), a randomized trial is a study that divides participants into 'experimental' (who will get, for example, vitamin C) and 'control' groups (who will, for example, be left untreated). Supporting the view that RCTs eliminate all confounders, Mike Clarke, a UK director of the Cochrane Collaboration[11], states:

> In a randomised trial, ... any differences in the outcomes of the patients in the groups being compared will be due to either the interventions they were allocated to receive, or the chance variations that will always exist between groups of people (Clarke 2004).

Although it is acknowledged that other study designs can offer evidential support, it is widely believed that these other forms of evidence are inferior. In fact the RCT (more specifically, as we will see, often the placebo controlled, double blind RCT) is often considered to be the gold standard of medical research and is placed at the pinnacle of the 'hierarchy of evidence'. In order to understand the arguments to come, the 'hierarchy of evidence', 'RCTs' must be defined.

2.1. Justifying the place of the RCT at the top of the Hierarchy of Evidence

Before examining the hierarchy and its justification in any detail a more detailed description of RCTs is required.

2.1.1. The randomized trial: a definition

A randomized trial, or RCT, is a medical experiment where patients (or, since they are, or at least should be, volunteers, 'participants') are *randomly* allocated to receive either the experimental (or 'test') treatment, or the control treatment. Random allocation is a process whereby "all participants have the same chance of being assigned to each of the study groups" (Jadad 1998`, p.2). A fair coin, random number tables, or

[11] The Cochrane Collaboration is a global network of volunteers who compile systematic reviews of (mostly) RCTs. They are considered the 'empirical research' arm of the EBM movement.

random generators on a computer can achieve randomization. Although there can, in principle, be more than one test group and more than one control group, I will limit my discussion to the simple case where there is a single experimental and a single control group. The control group can either be another treatment, a 'placebo', or 'no treatment'[12]. A 'placebo' is a treatment that is capable of making people believe it is, or could be, the experimental treatment when in fact it is not. A sugar pill that is indistinguishable to the senses from 'vitamin C' could be a placebo.

Randomization is either simple or restricted. Simple randomization is when a fair coin toss or some other random process determines the group to which a patient in a trial is assigned. *Restricted* randomization deviates from simple randomization and comes in several forms. In all cases the purpose of restricted randomization is to equalize known potential baseline differences. Block randomization, for instance, ensures that the test treatment and control groups are in the desired proportion[13]. Usually, there would be the same number of participants in the test group as there are in the control group. Other forms of restriction, including *stratification*, *weighted randomization*, and *minimization*[14] can be used with or without blocking. These

---

[12] To say that the control group was not given any treatment at all might be misleading. These 'no treatment' control groups face a double-edged sword. Either they are more or less left alone, in which case the investigators lose control over whether they choose to treat themselves with some other treatment of their own accord; or, they are closely monitored in which case the effects of being monitored (which could well be similar to expectation effects) could play a role.

[13] More specifically, block randomization randomizes n individuals into k treatments of block size m (the sample size must be divisible by the block size). For example, if there were 10 individuals, and two study groups, then we might use two blocks, TC, and CT, where TC meant "first test treatment, then control", and CT meant "first control, then test treatment". Then, a coin toss would indicate which block (TC or CT), and thus an order, in which the next two participants in the trial would be assigned to the study groups (Jadad 1998`, p. 5). Larger block sizes, and even randomly chosen block sizes, can also be used (Bland 2000).

[14] The different methods for restricting randomization may be used along with blocking and each other. *Stratified* randomization attempts to keep relevant characteristics, such as age, sex, or disease severity, equal across groups. In a stratified design, participants are organized into various *strata*, and the members of each strata are randomized into the various treatment arms. So, for example, there could be a strata of males. Then, this strata would be randomized to the various study groups. With *minimization* "the first participant is truly randomly allocated; for

methods all take baseline factors (such as sex) that are believed to affect the outcome, and equalises their distribution in the treatment and control groups. For simplicity, and because only simple randomization is compatible with the view that RCTs belong at the top of the hierarchy[15]. I will limit my discussion to simple randomization.

In addition to being restricted, randomization can be concealed. Concealed allocation occurs when the investigators in charge of allocating participants to experimental and control groups are not aware of *which* is the experimental or control group. For instance, they could assign participants to group 'A' or group 'B' and not know which is the experimental treatment. Concealed allocation does not have to be random. In order to satisfy the requirement of concealment, the only condition is that the investigators in charge of allocation be unaware of whether a participant is assigned to the experimental or control group.

The next feature of an RCT is that it is *controlled*[16]. A *controlled* study is one where the effect of the test treatment is compared with a different dose of the test treatment, no treatment, a placebo, or standard treatment. In a placebo controlled, or 'no treatment' controlled RCT, the test treatment is considered effective if it demonstrates that it is more effective than the placebo control or no treatment. An RCT that compares the test treatment with existing established treatment standard treatment (sometimes called 'active' treatment, although this is, strictly speaking a misnomer since placebos can be 'active' as well) can be conducted in two ways. First, the experimental treatment

---

each subsequent participant, the treatment allocation is identified which minimizes the imbalance between groups at that time. That allocation may then be used, or a choice may be made at random with a heavy weighting in favor of the intervention that would minimize imbalance (for example, with a probability of 0.8)" (N.I.H 2006). Then, there is *weighted* randomization. For whatever reason, it may be desirable to have more participants in one particular group. In these cases, the randomization can be weighted so that on average, more people are assigned to the group. For further discussion see (Bland 2000; Armitage, Berry, and Matthews 2002).

[15] Once we admit that restricting randomization is necessary the difference between randomized and non-randomized studies becomes vague since many of the strategies for restriction are equally available to non-randomized studies

[16] Sometimes the 'C' in 'RCT' refers to 'clinical' rather than 'controlled'. The term 'clinical' refers to a study of human subjects, or: "pertaining to or founded on observation and treatment of participants, as distinguished from theoretical or basic science" (N.I.H 2006).

will be considered effective if it demonstrates *superiority* to the active control; or, the experimental treatment will be considered effective if it demonstrates that it is at least (roughly) as good as the standard treatment, or better[17].

The final feature of an RCT is that it is a trial, which, in this sense, refers to the fact that it is experimental. A medical study is said to be experimental when the investigators are in control of administering the experimental intervention. Martin Bland, for example, defines experimental studies as follows: "In experimental studies, we do something, such as giving a drug, so that we can observe the result of our action" (Bland 2000`, p. 5). Similar definitions can be found elsewhere (Armitage, Berry, and Matthews 2002`, p. 6); (Jadad 1998`, p. 2). Sometimes allocation of participants to experimental and control groups is taken to be a necessary feature of experiments but I believe this is a mistake (see appendix to this chapter).

To sum up, an RCT is a medical experiment aiming to determine whether an intervention is effective, and where subjects are randomly allocated to treatment and control groups. With the definition of RCTs out of the way the arguments for their superiority over the other study designs in the hierarchy can be stated.

2.1.2. The accepted justification for the RCT's privileged place at the top of the hierarchy

The Evidence-Based Medicine movement, which arose in the early 1990's, rapidly expanded to play a central role in the medical (and later, social science) communities. The term 'evidence based medicine' first appeared in 1991 (Guyatt 1991), and the number of citations to EBM has risen exponentially. Within 6 years there were over 500

---

[17] More specifically, in a placebo controlled RCT using a classical hypothesis test, the test treatment is considered effective if the *null hypothesis* (the hypothesis that there is no difference) is rejected in favour of the hypothesis that the test treatment is more effective. In a RCT that compares test treatment with standard treatment, the test treatment is considered effective when the null hypothesis (this time the hypothesis that the test treatment is inferior to the standard treatment) is rejected in favour of the hypothesis that the experimental treatment is as effective, or more effective than, the 'active' control. The justification for the latter type of trial is that an experimental treatment could represent an advance without being more effective than the best existing treatment. For instance, the new treatment could have more tolerable side effects (at least for some people), or a lower cost.

new citations to EBM per year. The number of new citations seems to have tapered off at around 5000 per year at the time of writing. There are now several journals dedicated exclusively to EBM, and several other journals, including the BMJ, accept the EBM view. Then, EBM has colonized several other fields. It is not uncommon to hear of evidence based social science, evidence based policy, evidence based education, and even evidence based science.

A central tenet of the EBM view, indeed many would say *the* central tenet, is that randomized trials provide the best possible evidence for the claim that an intervention is effective. Other types of evidence, although they acknowledge are sometimes useful, provide inferior evidential support. They arrange the various types of evidence in a hierarchy.

Although there is more than one hierarchy, they all agree that RCTs (or collections of RCTs) belong at the top. The hierarchies are best explained with the use of a table (see table below).

1. **Table 2.1: Hierarchy of study types**

Level of Evidence	Type of Evidence
1	Systematic reviews[18] of RCTs and/or large-scale RCTs
2	RCTs
3	Observational studies (case-control and cohort studies)
4	Clinical expertise and 'pathophysiological' rationale

Hierarchies from the EBM textbook, the regulatory bodies in the United Kingdom, and the United States all share the structure of the hierarchy (Harbour 2008; Phillips et al. 2001; Canadian Task Force on the Periodic Health Examination 1979; US Preventive Services Task Force 1996). I will outline the justification for the hierarchy by noting how RCTs supposedly solve the problems with other types of evidence. I will focus, however, on the difference between RCTs and observational studies, leaving a description of 'pathophysiological reasoning' and 'clinical expertise', and systematic reviews to another study[19].

In an 'observational study', the intervention is not administered by the investigators, but rather the results observed (usually by examining hospital records): "In observational studies, aspects of an existing situation are observed, as in a survey or clinical case report" (Bland 2000`, p. 5). For example, if investigators examined the records of patients suffering from the common cold who had taken vitamin C, and compared them with the records of patients who had not, it would be an observational study. Observational studies can be *cohort* studies, *case control* studies, or, although this can be disputed, *historically controlled trials* (see appendix to this chapter). I will not describe each type of 'observational' study here, as it would take us far afield to do

[18] A systematic review is a collection of several RCTs testing the same intervention. In some cases, the similarity of the collection is sufficient that the results can be integrated for purposes of statistical analysis and the systematic review becomes a 'meta-analysis' (Green and Higgins 2005).

[19] My simplification is not without relevance to the justification for the other types of evidence in the hierarchy. The place of systematic reviews of RCTs is parasitic upon the alleged epistemic value of RCTs themselves. Then, as we will see, some of the arguments for the superiority of RCTs over observational studies can also be used to argue that RCTs are superior to 'pathophysiological' reasoning and clinical expertise.

so[20], but rather limit my remarks to a brief description of cohort studies. A cohort study is a study in which a defined group of people (the cohort) is followed over time. The outcomes of people in subsets of this cohort are compared, to examine people who were exposed or not exposed ... to a particular intervention or other factor of interest" (Green and Higgins 2005). For example, cohorts of people who took vitamin C and who did not take vitamin C could be followed up (i.e. examine medical records) to see if there were different recovery rates from the common cold. Cohort studies can be prospective or retrospective, but in both cases the cohorts can be taken from the same time period.

The alleged danger with observational studies is that they seem particularly vulnerable to both selection and performance bias. A cohort study, for example[21], that compares people who took vitamin C with those who did not, is comparing people who *chose* to take vitamin C with those who chose not to. People who choose to take vitamin C may be different than those who do not – they may be more optimistic, for example. Or, those who take vitamin C could live in areas where vitamin C is more readily available, and these areas could, on average, be less polluted. In short observational studies suffer from what is called 'self-selection' or 'patient preference bias', a type of selection bias resulting from the fact that people who 'select themselves', or 'prefer' a particular type of treatment could be different in relevant ways from those who 'select themselves' for no treatment or a different treatment. In addition to 'self-selection' of the patients themselves, there is the potential selection bias resulting from the fact that physicians who recommend certain treatments and treat people in a certain way could be unique. For instance, physicians who recommend vitamin C could also happen to be more empathetic than physicians who do not. The additional empathy, and not vitamin C, could affect the outcome.

At least in principle, historically controlled trials seem as vulnerable to selection bias, at least in the form of self-selection, as other observational studies – those who chose to be treated by the old treatment in the past may be different from those who choose to be treated by experimental treatment later. Historically controlled trials also suffer from the additional problem that people, diseases, and general protocols for ancillary care all change over time in ways that could be relevant to the outcome.

---

[20] Descriptions can be found elsewhere (Green and Higgins 2005).

[21] Parallel arguments apply to case-control studies and historically controlled trials, so for simplicity I will limit my discussion of observational studies to cohort studies.

In addition to selection bias, observational studies appear to be particularly vulnerable to performance bias. If, for example, the effects of a new expensive drug for depression that was widely advertised, were compared with the effects of taking nothing at all in a cohort study, we could not be sure that any difference between groups was due to the drug. The expectations induced by the advertising, could be, on average, greater in the group taking the drug than in the group not taking the drug. If these expectations affect depression, which is entirely possible, then even if the recovery rate were greater in the group taking the drug, we could not be sure that it was the drug itself or the increased expectations. Likewise, those taking the drug could have been seen by different physicians than those not taking a drug (if those taking nothing saw anyone at all). The different physicians could treat patients differently, and this differing treatment, and not the drug, could affect the outcome.

These problems with selection bias and performance bias are allegedly solved by RCTs.

2.1.3. The promise of randomization: elimination of selection bias

The real problem of selection bias are allegedly solved by *random allocation* of participants to experimental and control groups. The 'official' EBM bible, for example, states: "Randomization balances treatment groups for prognostic factors, even if we don't yet know enough about the target disorder to know what they all are" (Straus, Richardson, and Haynes 2005, p. 118). For example, if a group of people are randomized to receive vitamin C or placebo, it is alleged that the randomization equalizes all baseline factors.

The alleged superiority of randomization was apparently supported by empirical evidence in which several treatments that appeared effective based on the observational evidence, as no different, or worse than 'placebo' when examined in randomized trial. Perhaps the best-known example of this was when the prophylactic use of encanide and flecainide (antiarrythmic drugs) to reduce myocardial infarction (MI, or heart attack). In Worrall's words,

> Another example often cited ... concerns a phenomenon called ventricular ectopic beats. After a myocardial infarction, the heart remains electrically unstable and sometimes throws off characteristic beats. Those patients who exhibited these ventricular ectopic beats showed a greater incidence of subsequent cardiac arrest than those who did not exhibit them. It seemed to make good sense therefore to suppress the ventricular ectopic beats in the expectation that this would reduce the risk of cardiac arrest. The beats could

> be repressed fairly straightforwardly by administering substances like encainide or flecainide (also used as local anaesthetics). This became standard treatment but when a randomized trial [despite the objections of some clinical experts who claimed it was unethical to perform a placebo controlled trial] was performed it showed a *higher* rate of mortality from cardiac arrest amongst those treated for the suppression of the beats. And this treatment has now been abandoned (Worrall 2007b`, p. 4)

In addition to the above cases, a few other RCTs revealed that current practice was ineffective or harmful. Human growth hormone (HGH) was administered to intensive care unit (ICU) patients when they appeared hypercatabolic (experiencing breakdown of body tissue leading to weight loss and wasting). The rationale for administering HGH was quite clear – HGH promoted the growth that was being impaired. However, a large-scale randomized trial revealed that HGH increased mortality (Takala et al. 1999). In yet another example, increased oxygen delivery was attempted by transfusing blood to critically ill patients whose organs were failing. The increased oxygen flow to the organs was thought to prevent failure. Again, however, a placebo controlled trial revealed that this procedure increased rather than decreased mortality (Hayes et al. 1994). Then, in non-critically ill patients, hormone replacement therapy (HRT) was thought to prevent cardiac disease based on extensive observational evidence, but it was shown to increased cardiovascular morbidity in a subsequent placebo controlled randomized trial (Rossouw et al. 2002). Note that in the example Worrall cites (as well as the others), the RCT evidence overturned not only the observational evidence, but also pathophysiological reasoning (the rationale for the interventions was sound) and expert opinion.

2.1.4. Worrall's critique of the alleged advantages of randomization

In spite of its surface appeal, there is good reason to pause before swallowing the view that RCTs deserve their place at the top of the hierarchy of evidence. For one, many medical treatments in whose dramatic effectiveness we have the greatest confidence have never been tested in RCTs of any description. These treatments include Automatic External Defibrullation to start a stopped heart, tracheostomy to open a blocked air passage, the Heimlich manoeuvre to dislodge an obstruction in the breathing passage, rabies vaccines and penicillin for the treatment of pneumonia. Assuming that the current view is correct, then our confidence in these seemingly effective treatments

is exaggerated, perhaps irrational. If RCTs are so much better than other study designs, how come other study designs seem to produce these robust results?

In fact the view that randomization is an epistemic good, while supported by some philosophers (Papineau 1994; Pearl 2000) has long been challenged by Bayesian philosophers of science (Urbach 1985; Savage 1976; Lindley 1993, 1982; Howson and Urbach 1993), and more recently, Worrall (2002, 2007).

Then, the debate over whether RCTs provide the best evidence has also received considerable attention. This attention has come from philosophers[22], as well as others[23]. Often intertwined with the debate over the epistemological status of RCTs is the more general 'Evidence-Based Medicine' (EBM) position on evidence (which, as we shall see, involves conferring a very special epistemic status to RCTs), which has also been given some specific attention from philosophers[24]. Still others argue that there are no hard-and-fast rules of evidence that can be universally applied to the testing all treatments (Greenhalgh 2006; Enkin et al. 2005). Developing what can be interpreted as a variation on this theme, Cartwright recently argued that RCTs are not the only 'gold standard' of evidence (Cartwright 2007). Lastly, some critics claim that the current view, and EBM in particular is merely the latest of a series of paradigms, that is accepted to a large degree because its proponents are powerful and exclude other views from being discussed (Holmes et al. 2006).

The most sustained critique of the EBM view that randomized trials provide the best evidence in the philosophical literature, however, has come from a series of papers by John Worrall (2002; 2007a; 2007b), where he considers five (Bayesian and non-Bayesian) arguments for the potential benefits of randomization, and finds them, on the whole, wanting.

Worrall notes that the 'strongest' argument – "'strongest' in the sociological sense that it is the one that has convinced most people in medicine" (Worrall 2007b`, p. 18) – claim that randomization eliminates all confounders, both known and unknown. Worrall argues that this argument is patently wrong:

---

[22] See, for example (Worrall 2007b, 2007a, 2002; Cartwright 2007).

[23] See, for example example (Cochrane 1972; Barton 2000; Pocock and Elbourne 2000; Benson and Hartz 2000; Penston 2003; Brighton et al. 2003; Higgins 2005) .

[24] See, for example (Worrall 2002; Ashcroft and ter Meulen 2004; Ashcroft 2004; Borgerson 2005; Bluhm 2005; Grossman and MacKenzie 2005).

> Clearly the claim as made is quite trivially false: the experimental group contains Mrs Brown and not Mr Smith, whereas the control group contains Mr Smith and not Mrs Brown, *etc* (Worrall 2007b`, p.18)

Worrall's accusation is clearly justified. Flipping a coin to divide, say, two males and two females into two groups would not necessarily create groups with equal numbers of males and females. Even if the random process created two groups with equal numbers (that is, 2 people per group), there would only be a 2 in 3 chance of creating groups with an equal number of males and females. This problem arises whether or not the factor is known or unknown.

Of course in a larger trial, the empirical law of large numbers dictates that it is, in fact, unlikely that the proportion of common factors such as being male or female will differ widely. However large trials do not solve the problem for two reasons. First, there are, at least in principle, an infinite number of differences, and no matter how large the trial is we can never really be sure that many of the relative differences are unequally distributed. The second problem with large trials is contingent upon the logic of significance testing, which is currently taken for granted by the vast majority of the medical community. With classical hypothesis tests, the absolute difference between placebo (or no treatment) and experimental intervention required to conclude that the experimental treatment is 'effective' shrinks (roughly) in proportion with the number of people in the trial. This means that the amount by which the experimental treatment must be superior to placebo in order to 'prove'[25] that the experimental treatment is 'effective' decreases progressively as the size of the trial increases. Hence, the *decrease* in likelihood that the larger trial has gross baseline differences is coupled with an *increase* in ease to 'prove' effectiveness with a classical hypothesis test. For this reason, many authors insist on, in addition to the results of the classical hypothesis test, measures of effect size. The EBM movement, for example, recommends using the 'number needed to treat', or 'NNT', which is a measurement of how many people would have to be treated with the experimental intervention to get one positive outcome. An NNT of one is obviously ideal: we treat one person and get one positive outcome. An NNT of 100, on the other hand will, in most circumstances, be far less

[25] I am speaking very loosely. Obviously, in classical hypothesis testing nothing is proved, but rather the 'null' hypothesis (in this case that the experimental treatment is less effective than, or as effective as, the placebo) is 'rejected' at a certain significance level, usually 0.05 or 0.01.

impressive regardless of the result of the significance test. Unfortunately the results of significance tests are too often taken as sufficient.

Perhaps because the trivial falsity of the claim that randomization controls for all confounders, some texts that either pre-date the EBM movement or are not directly aligned with it admit that randomization alone does not control for *known* confounders. Bradford Hill, for example, states:

> When the results of a treatment are likely to vary between, say, the sexes or different age groups, then a further extension of this method may be made, to ensure a final equality of the total groups to be compared (Hill and Hill 1991, p. 220).

Other well-known medical textbooks make similar claims (Armitage, Berry, and Matthews 2002; Bland 2000). Even the EBM textbook takes this position implicitly when it recommends that investigators ask whether, after randomization, the comparison groups were "similar at the start of the trial" (Straus, Richardson, and Haynes 2005, p. 120). In short, the claim that randomization controls for known potential confounders is unsustainable and, upon close examination, not widely maintained.

However even those who recognize the need to control for known confounders claim that randomization eliminates *unknown* confounders. Fisher, for instance, claimed:

> [T]he full procedure of randomization [is the method] by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated [by being controlled]. (Fisher 1947, p. 19).

Fisher's claim about unknown confounders is echoed elsewhere. Bradford Hill states:

> We can equalise only for such features as we can measure or otherwise observe, but we also need unbiased allocation for all other features, some of which we may not even know exist. Only randomisation can give us that, and no form of equalisation can be a satisfactory substitute for it (Hill and Hill 1991, p. 221)

Simple random allocation, to Bradford Hill, is widely regarded as unwise when it is possible to consciously ensure that these potential confounders are distributed equally. However, the claim that unknown factors are equally distributed by random allocation is as unsustainable as the claim that known factors are equally distributed by random allocation.

Consider the example of randomizing four people once again. If we flip a coin to divide a group of four people (two male, two female) into two groups indefinitely, we will find that the most likely result is an equal distribution, but it does not follow that a *single* random division will produce equal groups. Recall the example of the two males and two females, call them M_1 M_2 and F_1 F_2. Now assume that a participant's sex is a known confounder, which is explicitly distributed evenly in two groups. Now imagine that whether a participant smokes is an unknown confounder, and that M_1 and F_1 are both smokers. If a stratified, blocked random design were used to divide the known confounders equally – say into M_1F_1 and M_2F_2 the unknown confounder will be unequally distributed.

Worrall then considers a *weaker* claim about the role of randomization in controlling for baseline differences, namely that randomization makes equal distribution of unknown factors *probable*. He states:

> A positive result in a randomized test, because the two groups are *probably* equal in all other respects, gives us, not of course foolproof, but still *the best* evidence of treatment effectiveness that we could possibly have. We do not eliminate entirely the possibility of 'bias' by randomizing, but we do 'eliminate' it 'in some probabilistic sense' (Worrall 2007b`, p. 20).

Worrall argues that this probabilistic argument only makes sense *in the long run*, and it does not help us determine whether, in one particular trial, some potential factor ('X') has been equally distributed:

> it can only amount to a claim about an *indefinite series of repetitions of the trial*: if you were to take a population (or perhaps a series of 'equivalent populations' whatever that exactly means), randomly divide it in two lots and lots of times and record the cumulative relative frequency of positive values of X in the two groups ... then in the indefinite long run that frequency would be the same in the experimental and control groups. ... But medical researchers involved in some particular trial do not make a random division indefinitely often, they only do it *once*! (Worrall 2007b`, p. 21).

Like the objection to the stronger claim, Worrall's objection to the claim that randomization makes equalization of baseline differences *probable* makes perfect sense. Randomizing two males and two females to two different groups might produce equal groups in the long run, but it obviously does not follow that the confounders will be equally distributed in a single trial.

In short, randomization neither controls for known nor unknown factors that could influence the outcome of the study, nor does it render likely that such confounders will be equally distributed. The best we can do is control explicitly for factors we know,

or strongly suspect, might affect the outcome of a study. Worrall concludes that "the argument that has convinced the great majority of the medical community that RCTs supply the 'gold standard' is without real foundation" (Worrall 2007b`, p. 21).

Another argument Worrall considers is that randomization underwrites the logic of classical significance testing. Returning to the quote from Fisher (1947, p. 19), it can be seen that:

> [T]he full procedure of randomization [is the method] by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated [by being controlled] (Fisher 1947, p. 19).

Because an in-depth discussion of the logic of classical significance testing, which would be required to analyze these arguments in any detail, would take us far afield, and because this subject has already received a considerable amount of attention, I shall restrict myself here to reporting what others, including Worrall (Worrall 2007b, 2007a) and among others(Savage 1976; Urbach 1985; Howson and Urbach 1993; Lindley 1993), have said.

First, as Howson and Urbach note, it is unclear that the argument about significance tests is justifiable on its own terms:

> Despite the widespread agreement that significance tests require randomization, expositions of the standard tests that are employed in the analysis of trials, such as the *t*-test, the chi-squared test, and the Wilcoxon Rank Sum test, barely allude to randomization (Howson and Urbach 1993`, p. 262-3).

Then, even if Fisher's argument were acceptable, if classical significance tests are unacceptable, it doesn't matter:

> as Bayesians have argued, even had it succeeded, the response might well have been 'who cares, given that in any case significance testing is so obviously and deeply problematic?'" (Worrall 2007b`, p.15-16).

A further argument Worrall considers is that randomization eliminates the attitudes of the allocators from introducing some factor that might influence the outcome:

> A third argument for the value of randomized control is altogether more down to earth. If the clinicians running a trial are allowed to determine the arm to which a particular patient is assigned then, whenever they have views about the comparative merits and comparative risks of the two treatments, there is some leeway for those clinicians to affect the outcome of the trial (Worrall 2007b`, p.21).

Worrall calls the potential bias introduced by the investigators (or 'clinicians' as he calls them) 'selection bias', although he acknowledges that the use of the term is ambiguous. As stated above, I will reserve the term 'selection bias' more generally to denote any difference between test and control groups, no matter how it is introduced (or created). I will use the term 'allocation bias' to denote bias introduced, either consciously or unconsciously by the investigators who allocate participants to treatment and control groups.

> Clinicians might, for example, - no doubt unconsciously – predominantly direct those patients they think are most likely to benefit to the new treatment or, in the other circumstances, they might predominantly direct those patients they fear may be especially badly affected by any side-effects of the new treatment to the control group (Worrall 2007b`, p.21).

The elimination of allocation bias is, as Worrall notes a "cast-iron argument for randomization" (Worrall 2007b`, p.22), but he qualifies this alleged benefit in two ways. First, he notes that it is the concealment of the allocation rather than the randomization itself which achieves, or at least is capable of eliminating allocation bias.

> Notice however that randomization as a way of controlling for selection bias is very much a means to an end, rather than an end in itself. The important methodological point is that control of which arm of the trial a particular patient ends up in is taken away from the experimenters - randomization (as normally performed) is simply one method of achieving this (Worrall 2007b`, p.22).

Any method of experimental allocation that is not in the hands of the investigators, whether it is random or not, will eliminate allocation bias. The second qualification Worrall makes is that, once suspected confounders have been controlled for explicitly, allocation bias will not introduce dramatic bias:

> Again to reiterate, those leading proponents of the virtues of randomization, Richard Doll and Richard Peto, acknowledge this point when writing that selection bias 'cannot plausibly give rise to a *tenfold* artefactual difference in disease outcome ...[but it may and often does] easily give rise to *twofold* artefactual differences. Such twofold biases are, however, of critical importance, since most of the really important therapeutic advances over the past decade or so have involved recognition that some particular treatment for some common condition yields a *moderate but important* improvement in the proportion of favourable outcomes' (Worrall 2007b).

In short, although randomization eliminates allocation bias, allocation bias is not the only source of selection bias, and may not be a dramatic confounder (at least if known confounders have been explicitly controlled for).

The fourth argument Worrall considers is that observational studies apparently exaggerate treatment effects:

> A fourth influential argument for the virtue of randomizing is that, no matter how the epistemological normative niceties about randomization play out, it is just *an empirical matter of fact* that other forms of trial have proved less reliable and shown themselves much more likely to produce a positive result that 'properly randomized' studies (Worrall 2007b`, p.22).

This fourth argument is supported by several 'meta-studies' that compared RCTs and (mostly) historically controlled trials from the 1970's and 1980's that tested the same intervention.

Worrall points to two problems with this argument. First,

> even if we accept the results of these meta-studies at face value, there is a clear circularity involved in the argument from those results to the claim that historically controlled trials 'routinely lead to false positive conclusions' (Worrall 2007b`, p.23).

In fact, in order to rule out the possibility that observational studies provide more accurate results, the possibility that RCTs underestimate treatment effects must be considered more seriously. Worrall cites a study by Black (1996) that argues this very point (Black 1996).

Second, the historically controlled studies with which the RCTs were compared in the earlier meta-studies may have been unrepresentative and particularly badly conducted. "Chalmers *et al* themselves suggest that the control and experimental groups in the historically controlled trials they investigated were patently 'maldistributed' with respect to a number of plausible prognostic factors" (Worrall 2007b`, p.24).

In fact, two recent studies (Benson and Hartz 2000; Concato, Shah, and Horwitz 2000) suggest that there is little difference between well-done observational studies and RCTs. Benson and Hartz (2000), for instance, conclude that there is

> little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials (1878).

Of course these later meta-studies have been criticized, notably by Pocock and Elbourne, mostly because the randomized and non-randomized studies chosen by these analysts may have been unrepresentative (Pocock and Elbourne 2000). Worrall objects that the only reason to supposed that the observational studies are unrepresentative is a prior commitment to the view that randomized trials provide the 'true' effectiveness. He states, "But again the only reason for thinking so seems to be a prior commitment to the idea that randomization (if properly done) is bound to be epistemically more telling"

(Worrall 2007b). In short, without the prior commitment to the view that randomized studies provide a 'true' measure of effectiveness, the argument that observational studies exaggerate effects is difficult to maintain.

The final argument Worrall considers for the advantage of randomization is that it apparently allows us to determine probabilistic causes[26]. As with the argument about randomization and classical tests of significance, a full review of this argument would take us far afield. Furthermore Worrall has already treated this argument in great detail elsewhere (Worrall 2007a). I will therefore limit myself to simply reporting what Worrall says:

> We all do, it seems, happily accept that there are true claims of the form X causes Y' where X and Y are generic events ('Smoking' and 'Lung Cancer' form a favorite example), and where the alleged connection fails to be deterministic. ... a's smoking tobacco (even heavily) does not inexorably bring about a's contracting lung cancer – yet we still want to say that smoking tobacco does cause lung cancer (Worrall 2007b`, p.26)

The non-deterministic claim about causality, Worrall notes, is not fully captured by the statement Prob(Y|X) > Prob(Y), i.e. Prob(lung cancer|smoking) > Prob(lung cancer). For instance, X and Y may be mere associations, or they may have a common cause. Smoking, for instance, may be the common cause of owning more ashtrays and of increased risk of lung cancer. So, Prob(lung cancer|ashtray ownership) > Prob (lung cancer), but ashtray ownership obviously does not cause lung cancer.

Attempts to develop accounts of probabilistic causality all incorporate some version of the common cause principle. Worrall notes that important contributors to the field of probabilistic causality, including Nancy Cartwright, David Papineau, and Judea Pearl[27]

> have explicitly claimed that it follows from their accounts that randomizing in a clinical trial is the vital ingredient in underwriting the claim that there is a *genuinely causal* connection between the treatment and the outcome, rather than a merely associational one (on the assumption, of course, that the outcome of the RCT is positive) (Worrall 2007b`, p.26).

But it is unclear how randomizing ensures genuine causality by ruling out chance associations or common causes. Imagine, for instance that we were testing the hypothesis that vitamin C caused speedier recovery from the common cold by

---

[26] Worrall treats this more fully in his *Why there's no Cause to Randomize* (2007).

[27] (Cartwright 1989; Papineau 1994; Pearl 2000)

randomizing people to get either vitamin C or placebo, and that we found the average recovery time in those who took vitamin C to be much shorter. But of course, being under 40 may also 'cause' speedier recovery from the common cold, and simple random allocation may well have resulted in more under-40s in the experimental group. Worrall concludes that the alleged connection between randomization and 'genuine' causality is merely the argument that randomization controls for all confounders, albeit dressed up differently:

> Cartwright, Papineau and Pearl are all in effect presenting the third argument for the special power of randomization considered above – that is, the argument that randomization controls for all possible confounders known and *unknown* – though they present it in somewhat different guises (Worrall 2007b`, p.27).

In short, the link between randomization and 'genuine' causes depends on the view that randomization rules out all confounders, whether they are known or unknown.

To sum up this section, Worrall is surely correct that the special role of randomization has been grossly exaggerated. In particular, the claim that randomization alone eliminates, or even reduces the probability of selection bias is patently false. Worrall, of course, does not claim that randomization does harm, but only that it is neither necessary nor sufficient for providing the best evidence.

There are, however, other aspects of RCTs and the hierarchy of evidence that are independent of the arguments for the alleged value of randomization, that have not received as much attention. These other potential features of RCTs are double blinding and placebo controls.

2.2. Double masking and RCTs: the reduction of performance bias

It is sometimes claimed that randomization controls for *all* confounders, whether they arise from selection *or* performance bias. Mike Clarke, the director of the UK Cochrane Collaboration, for example, states:

> In a randomised trial, ... any differences in the *outcomes* of the patients in the groups being compared will be due to either the interventions they were allocated to receive, or the chance variations that will always exist between groups of people (Clarke 2004`, italics added).

This claim seems to make no distinction between selection bias and performance bias. The arguments considered (and mostly rejected) by Worrall, focus on the claim that randomization eliminates selection bias. Yet on the face of it there is

nothing inherent about RCTs that controls for performance bias. It therefore seems that the scope of claim about randomized trials, even if we accept it, has been exaggerated.

Likewise, (as we saw above) the knowledge that a particular patient was getting the experimental drug could lead a dispensing physician to treat that patient differently from the patient getting nothing at all. To control for these confounders, 'double blinding' or 'double masking' is used. A double masked study is one whether neither the participants who receive the intervention nor those who administer the intervention are aware of who gets the experimental intervention. For example, in a double blind RCT of vitamin C versus sugar pill placebo, neither the participants nor the dispensing physicians know whether the particular pill they take (or administer) is vitamin C or a sugar pill. Not knowing whether they are in the experimental or control group, the participants' expectations, and any effects of these expectations, will not be confounded by the knowledge that they are taking a particular treatment. Similarly, if the physicians who administer the intervention do not know which participants are in the experimental or control groups, their treatment of all participants is will not be different because of their knowledge of who is taking the experimental treatment.

Because double masking seems to be able to rule out two potential confounders, double masking is regarded a methodological value. Indeed most evidence hierarchies do not place RCTs (or collections of RCTs) *simpliciter* at the top, but rather RCTs "with a very low risk of bias" (Harbour 2008). A low risk of bias, it is safe to assume, means, among other things (such as minimum size and concealed allocation), that the trials were double masked. Likewise, statements about the 'gold standard' of medical research do not refer to mere RCTs, but usually double blind, placebo controlled RCTs (Friedman 2004; Kaptchuk 2001).[28]

---

[28] An immediate reaction might be that concealed allocation, where neither the investigators in charge of allocating participants nor the participants are aware of which group they allocate participants *to*, achieves the aim of controlling for performance bias. However, this is not the case. Double masking is conceptually different from, and serves a different purpose from, concealed allocation. Concealed allocation is not necessary for blinding, and vice versa. To have concealed allocation without blinding, the allocation could be revealed once it had been complete. Or, to have blinding without concealed allocation, the investigators administering the intervention could be different from those in charge of allocation (this assumes, however, that

The view that double blinding is a methodological virtue has gone virtually unquestioned. The official EBM textbook, for example, states

> Blinding is necessary to avoid patients' reporting of symptoms or their adherence to treatment being affected by hunches about whether the treatment is effective. Similarly, blinding prevents the report or interpretation of symptoms from being affected by the clinician's or outcomes assessor's suspicions about the effectiveness of the study intervention (Straus, Richardson, and Haynes 2005, p.122).

The same text list the following questions that we should ask when appraising a study.

1. *Was the assignment of patients to treatment randomized?*
2. *Was the randomization concealed?*
3. *Were the groups similar at the start of the trial?*
4. *Was the follow-up of patients sufficiently long and complete?*
5. *Were all patients analyzed in the groups to which they were randomized?*
6. *Were patients, clinicians, and study personnel kept blind to treatment?*
7. *Were the groups treated equally, apart from the experimental therapy?*

(Straus, Richardson, and Haynes 2005`, p.117-123)

The steps for appraisal make it clear that double blinding is a feature of well-conducted trials, and therefore of what they take to be the best evidence. Clearly 2 already requires that the patients be blind; while 6 is an explicit repetition of this plus a requirement that trials be double (clinicians) and even triple ('study personnel') blind; finally 7, given the possibility of placebo effects, also clearly requires (double) blinding.

For good reason (at least on the face of it), the praise for double masking is not limited to EBM proponents. The United States Food and Drug Administration (FDA 1998), other authorities (Moher, Schulz, and Altman 2001) as well as prominent medical researchers and statisticians (Armitage, Berry, and Matthews 2002, p. 605; Bland 2000, p. 19; Hill and Hill 1991, p. 214; Greenhalgh 2006, p. 66; Jadad 1998, p. 20) all explicitly claim that double blinding is an methodological virtue. Alejandro Jadad devised a widely used scale (called the 'Jadad Scale') for appraising study quality. The scale allocates 1 (out of a total of 5) points for whether the trial is double-blind, and another point for describing the blinding adequately[29].

---

the allocation was concealed from the participants). To be sure, however, concealed allocation is the easiest way to achieve blinding and indeed there will usually be no reason *not* to conceal the allocation.

[29] The questions are, (1) Is the study randomized? (2) Is the study double blinded? (3) Is there a description of withdrawals? (4) Is the randomization adequately described? (5) Is the blindness adequately described?

In spite of its intuitive appeal and widespread support, there are several reasons to question double masking as a universal methodological virtue. For one, certain treatments resist being tested in double blind conditions. As noted last earlier, the Phillip's Paradox suggests that any treatment that turns out to have dramatic effects will not remain double blind[30]. Rather less dramatically, any control treatment that does not successfully imitate the appearance, taste, smell, and even side effects of the experimental treatment will be incapable of keeping a trial double masked. The requirement of informed consent[31], which I will take for granted, makes imitation difficult for many treatments. For instance, a trial comparing exercise with a placebo pill (or indeed anything other than a similar programme of exercise) will be impossible to keep double masked. The participants will know that they are enrolled in a trial of exercise versus placebo pill, so if they receive a pill they will know they are in the placebo group and their expectations regarding recovery could be lower.

It seems strange – to say the very least – that an account of evidence should deliver a purely *a priori* judgment that a certain type of claim can never be supported by 'best evidence'. It would of course be different if the claims at issue were pseudoscientific – untestable. But so far as treatments with large effects go at least, the claim that they are effective is highly testable and intuitively it would seem that they should receive much greater support from the evidence than do claims about treatments with only moderate effect sizes. Hence the claim that double blinding is a universal virtue is arguably inconsistent with educated scientific common sense.

---

[30] There are further difficulties with keeping a trial double blind, which I shall discuss below and in chapter 5.

[31] The requirement of informed consent can be traced to the Nuremburg code, and was designed to prevent a repeat of the unethical trials conducted on human subjects by the Nazi's in World War II. The code states that the research subject

should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonable to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment (N.I.H 2004).

The third reason to question the value of double masking is their tendency to reduce external validity. A trial is *internally valid* " to the extent that its outcome truly measures the impact of the treatment on the outcome *so far as the trial population is concerned* (Worrall 2007b`, p. 10). A trial is *externally* valid to the extent that its results are applicable to the wider public ('target population') in conditions of routine practice.

In a placebo controlled trial where the aim is to test whether the experimental treatment is superior to the control treatment, internal validity refers to the justification for the inference that apparent superiority of the experimental treatment is due to the actual superiority of the experimental treatment *for the trial population.* In an 'active' controlled trial where the aim is to test whether the experimental treatment is at least (roughly) equivalent to the control treatment, internal validity refers to the justification for the inference that apparent equivalence or superiority (rather than inferiority) of the experimental treatment is due to the actual similar or superior effectiveness of the experimental treatment *for the trial population.* In both cases internal validity refers to the grounds for supposing that the inference from the observed outcome of the study to the effectiveness of the test treatment *for the participants in the trial (the 'trial population')* is valid. Since double masking appears to rule out one possible explanation for the observed outcome other than the characteristic feature of the test treatment, namely participant expectations and dispensing dispenser bias, it would appear as that double masking increases the internal validity of the study.

Even if it is true that double masked trials generally have higher internal validity than open (not masked) trials (which is still an open question at least in some cases), the reverse may be true if we consider *external* validity. It is important to note that external validity is not merely "a reflection of the general inductive scepticism inherent in 'Hume's Problem" (Worrall 2007b`, p. 11). The population of volunteers in the trial is rarely, if ever, representative of the population who could benefit from the experimental treatment if it is approved for marketing (the 'target population'). The worry about external validity is also more than a 'philosophical quibble'. It has been found, for example, that even the most effective antidepressants for adults are not effective for children (Bylund and Reed 2007; Deupree, Reed, and Bylund 2007) In another example cited by Worrall, the drug benoxaprofen (trade name Opren):

> a nonsteroidal antiflammatory treatment for arthritis and musculo-skeletal pain. This passed RCTs (explicitly restricted to 18 to 65 year olds) with flying colours. It is however a fact that musculo-skeletal pain predominantly afflicts the elderly. It turned out that when the '(on average older) target

> population' were given Opren there was a significant number of deaths from hepato-renal failure and the drug was withdrawn.) (Worrall 2007b`, p. 10).

In fact, often up to 70% of potentially eligible participants for a trial are excluded according to poorly reported and even haphazard criteria (Penston 2003; Van Spall et al. 2007). Also, it must be remembered that only a fraction of the target population are potentially eligible for trials (Travers et al. 2007). Then, addition to differences between trial and target populations, differences between trial conditions and conditions in routine practice can also threaten external validity. If the trial, for example, includes intense follow up of patients to ensure that they follow a difficult, treatment regime, it may be unreasonable to suppose that the regime would be followed in routine practice where there is neither the time nor resources to follow patients so closely. It is hardly novel to suspect that the effects of a treatment in tightly regulated trial conditions and for highly selected trial participants might not have the same effects (or might have worse side effects) for the 'target' population.

Double masking, of course, adds a further difference to trial conditions. In routine practice, people are usually given an option of which treatment they prefer, and they are treated openly. These differences between trial and routine practice conditions could, at least in principle, threaten the external validity of the trial.

In short, the apparent value of double masking is clear and surely justified in many cases. Yet at the same time it is clearly not of great value in other cases, the 'Phillips Paradox' results from viewing double masking as a virtue of clinical trials of dramatically effective treatments. Further, it seems strange to deliver *a priori* judgements that certain treatments (including those that are dramatically effective) cannot be supported by 'best' evidence. In the coming chapters, I will question the value of double masking, and, using 'scientific common sense' as a standard, attempt to discern when double masking is a virtue and when it is not.

Another potential feature of trials that is not addressed by the arguments for randomization, namely the use of placebo versus 'active' controls.

2.3. The Alleged Superiority of Placebo over Active Controls

To achieve double masking, it must be possible to deceive both participants and investigators into believing that the control intervention could be the experimental intervention and vice-versa. For this to happen the control group cannot be withheld treatment altogether. A participant in a trial who gets nothing at all will obviously know

(given the requirement of informed consent) that she is not getting the experimental intervention. Hence, the participants in the control group are usually given 'placebos' or an existing treatment..

Although, for the purposes of double blinding either 'active' controls (an established, existing treatment for the same disorder) or placebos will do, the use of active controls is supposedly problematic for at least two important methodological[32] reasons. Thus, it is sometimes claimed that RCTs, in order to be considered of the highest quality, must employ placebo controls. Kaptchuk, for example, states that "A placebo controlled RCT is considered medicine's most reliable method" (Kaptchuk 2001, p. 541). Temple and Ellenberg, for example, state:

> Unfortunately, ACETs [active controlled trials] are often uninformative. They can neither demonstrate the effectiveness of a new agent nor provide a valid comparison to control therapy unless assay sensitivity can be assured, which often cannot be accomplished without inclusion of a concurrent placebo control group (Temple and Ellenberg 2000).

Simliar remarks can be found elsewhere (ICH 2000; Gomberg-Maitland, Frison, and Halperin 2003; Hwang and Morikawa 1999).

First, placebo controlled trials, but not 'active' controlled trials, possess *assay sensitivity*. The sound idea behind the assay sensitivity argument is that, if an experimental treatment proves equivalent or superior to an 'active' control, we cannot be sure that the test treatment is effective. The active control could be ineffective or worse, harmful. For example, it would be unwise to conclude from the fact that bloodletting (the experimental treatment) was at least as effective as lobotomies for psychosis, that bloodletting was effective. Both treatments could be no better (or indeed, far worse!) than doing nothing at all. Indeed many have claimed that before the era of modern scientific medicine, that most treatments used were mere placebos or worse. Grünbaum, for instance, states: "History teaches us that many well-intentioned treatments were *worse than useless*" (Grünbaum 1986`, p. 28).

Second, placebo controlled trials, but not 'active' controlled trials, supposedly provide an *absolute measure of effect size*. The outcome in the experimental group is the total average effect of the treatment, which is composed of the effects of the

---

[32] It is also worthwhile noting that placebo controls are ethically and practically problematic. I discuss these issues briefly in chapter 6.

characteristic features plus all other features (such as expectation effects). For instance, the outcome in the vitamin C group is due to the effects of the characteristic features of vitamin C (i.e. ascorbic acid) in addition to the effects of being prescribed a pill by an authority figure in a white suit, expectation effects, etc. The outcome in the placebo control group is (or should be) the total average effect *less* the effects of the characteristic features. A placebo vitamin C, for instance, is prescribed by the same (or a similar) authority figure in a white suit, and (if the trial is double blind) induces similar effects. Described this way it is easy to see how subtracting the average effects of the placebo group from the average effects of the control group could, at least in principle, provide an average measure of the 'absolute' effect of the characteristic treatment features.

These alleged *methodological* advantages of placebo controlled trials are touted by the International Conference on Harmonization (ICH) E10 document, produced and endorsed by the regulatory bodies of the United States, Japan, and the European Union (ICH 2000), as well as others (Senn 2005; Temple and Ellenberg 2000; Ellenberg and Temple 2000).

A few notable exceptions notwithstanding (Anderson 2006; Kirsch 2000), these alleged methodological advantages of placebo controlled trials have gone unchallenged. However, in spite of what seem to be good reasons to prefer placebo controls (at least in some cases), there are good grounds to question the view that placebo controls are superior to active controls.

To begin, some treatments resist being imitated by placebo controls for methodological reasons alone. With treatments such as vitamin C pills, and indeed most pills, a placebo control that a patient could mistake for real vitamin C is relatively straightforward to design. We simply remove the characteristic ingredient and replace it with an ineffective substance such as sugar. Metaphorically, the superficial appearance, or the 'shell' is kept the same while changing the 'characteristic feature' or 'filling'[33]. With exercise this straightforward technique for designing placebo controls is not

---

[33] The design of placebo controls even for pharmaceutical drugs is more complicated that I have indicated here. For one, the coating is not usually wax, and for another the coating 'contains' more than the characteristic ingredient. It will sometimes contain other agents to aid in the digestion of the drug, etc. Still the process for designing a placebo pill is *relatively* straightforward, at least at a first glance.

possible. For one, deciding what the characteristic features of exercise are – they could include sweating, increased heart rate, and muscle fatigue – is problematic. But even if we could decide for sure that any therapeutic effects of exercise are due to increased heart rate and muscle fatigue, it would be difficult to find an intervention that is superficially similar to exercise, yet does not involve increased heart rate or muscle fatigue. In particular, it seems difficult to design a treatment that is capable of making people think they are doing exercise without actually doing exercise[34]. Exercise is not the only treatment for which it is problematic to design placebo controls. Other treatments include acupuncture, some surgical techniques, and any kind of 'talking' therapy (such as psychotherapy, cognitive behaviour therapy).

As a result, these treatments are banished to being supported by (at best) second best evidence. Again the complaint here is not to deny that it may be true that evidence for the effectiveness of exercise for some complaint, say depression, is as a matter of fact weak; the complaint is the *a priori* nature of the ruling that this must inevitably be so. This view means that interventions which are very difficult to disguise, and hence for which plausible placebo controls are difficult to produce – one such example is exercise as a possible therapy for depression – are *automatically* barred from being supported by best evidence. Of course it may (or may not) be true that exercise is therapeutic for depression. It is a contingent matter whether or not we have any evidence for its effectiveness, but it surely seems an unattractive consequence of a methodology if it pronounces in advance, and independently of any facts about the world, that particular types of treatment are forever banned from being supported by 'best evidence'.

The unfortunate consequences of the view that certain treatments are excluded from support by the 'best' evidence are difficult to understate. When policy makers are deciding which treatments to recommend as a first line of attack, their decisions will (and should) be based on which treatments are supported by what they take to be the

---

[34] Current clinical trials of exercise use relaxation or supervised flexibility as 'placebo' controls for exercise (McCann and Holmes 1984; Dunn et al. 2002, 2005). I will argue in chapter 4 that these treatments may not be 'legitimate' placebo controls. For the moment note that, *prima facie*, there is a definite difference between a placebo drug pill, where the characteristic chemical is removed and replaced with some ineffective substance such as sugar, and conceivable placebo controls for exercise.

strongest evidence. But if they accept the view that double blind placebo controlled trials are the best possible study type, regardless of the objective effectiveness of potential treatments, drugs with moderate effects will tend to be supported by (what is taken to be) the best evidence, leaving other, 'complex' or dramatically effective treatments supported by lesser evidence[35].

In sum, there seem to be good reasons to prefer placebo controlled trials and equally good reasons, both methodological as well as ethical and practical, to question the apparent methodological superiority of placebo controlled trials. In the coming chapters I will examine the relative benefits and risks of placebo and 'active' controls and argue that, more often than is commonly believed, 'active' controlled trials are as methodologically sound as placebo controlled trials.

2.4. Appendix: Are Observational Studies Experiments?

According to the accepted view, in order to qualify as a clinical 'experiment', the investigators must be in control of administering the intervention, and (sometimes) of allocating participants to experimental and control groups. 'Historically controlled trials' are the type of study Worrall (2002, 2007b) contrasts with RCTs seem to satisfy the first, but not the second condition for being considered experimental. In these studies a new intervention is introduced experimentally, and its effects are compared with a historical control, namely past effects of standard intervention or no intervention. Historically controlled studies seem to satisfy one of the conditions for being considered 'experimental' studies since the experimental intervention is administered by the investigators. However, the investigators in a historically controlled study are not generally in control of what is given (or withheld) to the control group.

It could be disputed that a necessary feature of an experiment is that the allocation must be 'artificial' – that is, achieved by the investigators. In fact, the view that allocation must be artificial is not often made explicit, but the term 'non-randomized' is often identified with 'non-experimental'. In a paper criticising the view that RCTs always provide the best evidence, John Concato identifies RCTs with experiments. In the text he identifies RCTs with experimental studies, and it is quite clear that he views observational studies as non-experimental studies. The title to a

---

[35] The fact that patentable drugs are generally more expensive than other treatments makes matters all the worse.

section of his article is: "A more balanced view of observational and experimental evidence" (Concato 2004). Lamenting this state of affairs, Worrall notes "the RCT is often unthinkingly identified with 'the experimental method'" (Worrall 2007b`, p. 22).

The first condition for being considered experimental, that the investigators be in charge of administering the intervention, can also be disputed. In fact, this view leads to the view that 'experimental' methods are not available to astronomers, geologists, or geographers, which is surely incorrect. Indeed Eddington's solar eclipse experiment, on this view, would not be an 'experiment', since the investigators did not manipulate the celestial bodies they measured.

Mill, drawing what I believe is a more accurate distinction, separated 'natural' from 'artificial' experiments to distinguish between studies where the investigator was in control of administering the intervention, and those where the investigator was not. Mill would classify observational studies as natural experiments and RCTs as 'artificial experiments', implying that the scientific, experimental method was available to both. However because of the widespread use of the term 'observational study', and because it does not play a central role in the body of my thesis, I will continue to use the term 'observational study' to denote studies where the investigators are neither in charge of allocating participants nor administering the experimental. From this point of view, historically controlled trials are observational studies.

3. Chapter Three. Evidence from a more fundamental viewpoint

Scientific knowledge is merely a development of ordinary knowledge or common-sense knowledge

- (Popper 1969)

...when you have eliminated the impossible, whatever remains, however improbable, must be the truth

- (Conan Doyle 1890)

In making chemical experiments, we do not think it necessary to note the position of the planets; because experience has shown, as a very superficial experience is sufficient to show, that in such cases that circumstance is not material to the result

- John Stuart Mill (Mill 1843[1973])

3.1. The Bedrock of Scientific Method: Common Sense

The double-blind, randomized trial is considered to be the 'gold standard' of evidence in medicine. Yet ironically, interventions most would consider to be most strongly supported by evidence, including drinking water to recover from severe dehydration, the Heimlich manoeuvre to dislodge an obstruction to the airway, and using a defibrillator to start a stopped heart, have never been tested in randomized trials of any description. The problem with the view that randomized trials belong at the top of the 'Evidence-Based Medicine' (EBM) hierarchy, I will argue, is that the view is not supplemented with a commitment to the view that the best evidence counts against more rival hypotheses than other evidence. Because, as we shall see, this view is widely supported, I will call it 'scientific common sense'. Although randomized trials generally counts against more plausible rival hypotheses than other evidence, in many cases other types of evidence rule out equally many.

Suppose, for example, a study compares healthy people who take vitamin C with unhealthy people who do not, and the trial result was that those who took vitamin C overcame the symptoms of the common cold more quickly. Clearly 'scientific common sense' would rule that this was *not* evidence that provided strong support for the experimental hypothesis. Although the evidence would be explained by the hypothesis that vitamin C causes cold symptoms to disappear more quickly, it does count against the, extremely plausible, *rival* hypothesis that people who are generally healthier overcome colds more quickly.

There have been many attempts to develop an account of how evidence should be taken to impinge on theories. Indeed this problem is perhaps *the* central problem of the whole of the philosophy of science. These attempts have all differed from one another in detail and have led to a number of interesting, though often abstruse, debates. However, so far as recent issues about evidence in medicine are concerned, my belief is that we can go very far on the basis simply of judgments on which pretty well every serious philosopher of science will agree, namely that strong evidence both supports the experimental hypothesis and counts against plausible rival hypotheses.

I take the 'scientific common sense' intuition to be uncontroversial: if there is a rival that remains highly plausible in light of the evidence, then, *ceteris paribus*, it would be irrational to take the evidence as support for the experimental hypothesis. In fact, as I will contend, all well worked-out accounts of scientific method embody this intuition in some way or other. I use the fact that accounts of method as outwardly distinct as those offered by Popper, Mill, and Bayesians (as well as many others) all converge in support of 'scientific common sense' to support my argument. I will then suggest that whenever a more formal analysis leads to a conclusion – such as that there is no telling evidence except from a randomized trial – which sometimes conflicts with 'scientific common sense', then we should question the more formal analysis. First, however, I must say a few words about the skeptical objection that there are, at least in principle, an infinite number of rival hypotheses.

3.2. The Limits of the Skeptical Objection in the Context of the Philosophy of Science

In order to rule out plausible rivals, the rivals must be identified. Yet the thesis of underdetermination reminds us that there are, at least in principle, an infinite number of hypotheses that can explain any evidence. Similarly, the Duhem/Quine thesis reminds us that evidence never rules out an isolated hypothesis, but rather a conjunction of hypotheses. I will deal with each of these problems in turn.

The problem that there are, in principle, an infinite number of rival hypotheses seems to present an insurmountable block for 'scientific common sense': if there are an infinite number of rivals, then we obviously cannot rule them out. I will reply to this objection by putting it into the framework of the relevant alternatives theory (RAT) in classical epistemology. The RAT states that knowing a true proposition requires being able to rule out relevant alternatives to that proposition (Dretske 1970, 1981; Lewis

1996). The problem with the RAT, is that it turns out to be quite difficult to identify and rule out relevant alternatives. This problem is, of course, quite similar to the problem with 'scientific common sense'.

Various attempts have, of course, been made to limit the class of relevant alternatives. Dretske (1981), for example, claims that an alternative can simply be "*too remote* to qualify as relevant" (p. 376). Lewis offers a more sophisticated account and claims that the relevance of an alternative is determined contextually, and he provides seven rules of relevance which clarify how context determines relevance (Lewis 1996; Shaffer 2001).

I will not examine these arguments in any detail here, but simply note that it is unclear whether, in classical epistemology, they are sufficiently powerful to withstand the radical skeptic. For starters, the skeptic can always insist that the evil demon alternative is possible. RA theorists such as Dretske (1981) seem to acknowledge this failing:

> There are always, it seems, possibilities that our evidence is powerless to eliminate, possibilities which, until eliminated, block the road to knowledge. For if knowledge, being an absolute concept, requires the elimination of *all* competing possibilities (possibilities that contrast with what is known), then, clearly we seldom, if ever, satisfy the conditions for applying the concept. (p. 365).

The possibilities Dretske refers to no doubt include the possibility that an evil demon is deceiving us. For the purposes of this work I will take it for granted that there is no generally accepted response to the evil demon objection.

In science, however, the class of plausible rivals ('relevant alternatives') is more clearly defined, and is provided by background knowledge. For instance, the background knowledge at the time that Einstein proposed the theory of relativity included the hypothesis (which itself was supported by evidence) that Newton's Laws could accurately describe and predict the motion of heavenly bodies. Hence, 'scientific common sense' dictates that good evidence must both support Einstein's theory and count against Newton's theory. Eddington's famous observation of the deflection of light from Mercury during the solar eclipse of 1919 provided evidence which satisfied this condition and is regarded as good evidence that Einstein's theory rather than Newton's was the more acceptable.

Virtually all serious philosophers of science have acknowledged the role of background knowledge. Hempel described the role of background knowledge as follows:

> ... the credibility of a hypothesis *H* at a given time depends, strictly speaking, on the relevant parts of the total scientific knowledge at that time, including all the evidence relevant to the hypothesis and *all the hypotheses and theories then accepted* (Hempel 1966, p. 45, italics added).

Howson describes the importance of background knowledge in the following passage:

> A large book found in the street is by itself not evidence that Jones killed Smith, but given the further information that Smith was killed by a blow to the head with a large object, in that particular street, and that the book was damaged and had Smith's blood and Jones's fingerprints on it, it is. Evidence issues in the enhancement or diminution of the credibility of a hypothesis, and this capacity will be determined only in the context of some specified ambient body of information (Howson 2000, p. 179).

Although Mill does not mention background knowledge explicitly, he is clear that certain rival hypotheses need not be considered:

> In making chemical experiments, we do not think it necessary to note the position of the planets; because experience has shown, as a very superficial experience is sufficient to show, that in such cases that circumstance is not material to the result (Mill 1843[1973], I.III.vii).

This implies that certain facts are material to the result and therefore must be controlled for.

With regards to the problem of eliminating all possible alternative explanations, Howson is convinced that it is fundamentally unsolveable. When criticizing the view that *randomization* eliminates all rival hypotheses, he states "It is, of course, the problem of alternative explanations all over again. Randomization does not solve it. Nothing does" (Howson 2000, p. 51).

An objector might persist and insist that we need not be radical sceptics looking for evil demons to undermine 'scientific common sense'. We might grant that background knowledge specifies a class of relevant alternatives, but point out that our background knowledge could be mistaken. If so, then our evidence might appear strong because it ruled out the putative rivals, but the putative rivals might be mistaken and, moreover, there might be relevant alternatives that background knowledge failed to pick out.

The strength of this objection, however, depends (in a sense) on the view that our evidence must provide evidence for the *truth* of a hypothesis. In this thesis I will take for granted that scientific knowledge is fallible. In Popper's words:

> The fact that, as a rule, we are at any given moment taking a vast amount of traditional knowledge for granted ... creates no difficulty for the falsificationist or fallibilist. For he does not *accept* this background knowledge; nether as established nor as fairly certain, nor yet as probably He knows that even its tentative acceptance is risky, and stresses that every bit of it is open to criticism, even though only in a piecemeal way. We can never be certain that we shall challenge the right bit; but since our quest is not fore certainty, this does not matter (1969, 10.4.xvi).

The fact that background knowledge is fallible means that all hypotheses are tentatively supported, and that the strength of evidence will be relative to the state of our background knowledge. Popper makes this point in the following way:

> The old scientific ideal of *episteme* – of absolute, demonstrable knowledge – has proved to be an idol. The demand for scientific objectivity makes it inevitable that every scientific statement must remain *tentative for ever*. It may be corroborated, but every corroboration is relative to other statements which, again, are tentative. Only in our subjective experiences of conviction, in our subjective faith, can be 'absolutely certain'" (Popper 1968`, section 85).

Popper is not alone in his view that scientific knowledge is fallible and tentative. The claim that medical knowledge as, at best, tentative, has also been expressed by some of the most prominent medical statisticians. Sir Austin Bradford-Hill, for example, states:

> All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time (Hill and Hill 1991).

Martin Bland asserts, "No piece of research can be perfect and there will always be something which, with hindsight, we would have changed" (Bland 2000`, p.2).

In short, our ability to identify *all* rivals is necessarily limited. At the same time, the fallibility of scientific knowledge does not imply that our efforts to identify and rule out plausible rivals are futile. Background knowledge provides us with information about which rivals are plausible – factors are worthwhile controlling for.

My response to the objection, due to the Duhem/Quine thesis, that it is impossible to rule out an isolated hypothesis, or indeed an isolated subset of plausible rivals, is also pragmatic. In order to test a hypothesis, we must take a large amount of 'background knowledge' for granted. Popper (1968) for example, described the necessity of background knowledge in the following manner:

> While discussing a problem we always accept (if only temporarily) all kinds of things as *unproblematic*: they constitute for the time being, and for the discussion of this particular problem, what I call our *background*

> *knowledge*. Few parts of this background knowledge will appear to us in all contests as absolutely unproblematic, and any particular part of it *may* be challenged at any time (10.4.iv).

In short, although it is, strictly speaking, impossible to identify all rival hypotheses or even rule them out, 'background knowledge' in medical science will provide us, tentatively, with both a class of rivals that are deemed plausible, and with a class of hypotheses that can be safely (again tentatively) left unquestioned.

3.3. Popper, Bayes, and Mill: All endorse 'scientific commonsense'

Popper, Mill and the Bayesians all differ in any number of ways that philosophers of science have identified and debated. But I will show that they all agree on the 'commonsense' principle of evidence that I just articulated – and I take it that this agreement is powerful support for the idea that this principle, whatever its ultimate justification, is indeed part of educated or 'scientific common sense'.

3.3.1. Bayes and ruling out rival hypotheses

The Bayesian model of inductive inference uses Bayes' Theorem to tell us what the probability of a hypothesis is, given some evidence. There are many variations of Bayes' Theorem as well as descriptions of Bayesian confirmation theory. In the subsequent discussion I will follow Howson's (2000, p. 178-183) exceptionally clear exposition. Using a particular formulation of Bayes theorem, the probability of a hypothesis given some evidence (the 'posterior probability') can be represented as follows:

$$p(h \mid e) = \frac{p(h)}{p(h) + fp(\neg h)}$$

Where $p(h)$ is the probability of the hypothesis, and f is the 'Bayes factor', or, more precisely the 'Bayes factor in favour of $\neg h$ over h:

$$\frac{p(e \mid \neg h)}{p(e \mid h)}$$

It is clear that the posterior probability increases as the Bayes factor in favour of not-h decreases. This means that, other things being equal, the less probable the

evidence would have been were the hypothesis false[36], the greater the posterior probability. If the evidence were just as likely had the hypothesis been true as it would be if the hypothesis were false (and therefore the Bayes factor = 1), then the evidence does not add any credibility to the hypothesis. In this case the posterior probability of the hypothesis would be the same as the prior probability of the hypothesis[37]. On the other hand, if the evidence were *impossible* if the hypothesis were false (and therefore $p(e|\neg h) = 0$), then the posterior probability of the hypothesis will be 1 (this is clear if we manipulate Bayes' theorem substituting '0' for 'f').

In practice, however, the Bayes factor is neither 0 nor 1, but something in between. The farther from unity the Bayes factor is, the more informative is the evidence. This means that good evidence in favour of the test hypothesis will involve a situation where the Bayes factor in favour of $p(e|\neg h)$ is as low as possible.

The question is, how do we design an experiment – produce evidence – such that $p(e|\neg h)$ is very low? Intuitively speaking, we would attempt to eliminate every possible factor that might cause e to arise other than h in advance by the experimental design. As mentioned earlier, strictly speaking we can't possibly eliminate all the possible causes of e. But we can set out to minimize the influence of the ones we do know about.

More formally, if we can limit (using background knowledge), the class of (mutually exclusive) probable rivals to h, call them $\{h_i\}$, then the probability of the evidence given that the hypothesis were *false*, $p(e|\neg h)$, is equal to the sum of the factors $p(e|h_i)\cdot p(h_i)$. (We are restricted, of course, to cases where $p(h_i) > 0$.) That is,

$$p(e \mid \neg h) = \sum p(e \mid h_i) \cdot p(h_i)$$

As we reduce the probability of the evidence given the rivals $\{p(e|h_i)\}$, and/or the prior probability of the rivals $\{p(h_i)\}$, the posterior probability of the hypothesis rises. In Howson's words,

> ... given the body of background information we have accepted which describes the class of plausible alternative explanations of a positive result, designing a test so that the result if it occurs cannot plausibly be attributed

---

[36] And, of course, where $p(e|h)$ is high. (In many cases the hypothesis entails the evidence, in which case $p(e|h)$ will be equal to 1.) I will refrain from repeating this obvious fact for simplicity.

[37] This is clear if we manipulate Bayes' theorem algebraically substituting '1' for 'f', and '1 – $p(h)$' for '$p(\neg h)$'.

> to any of them *just* is to design a test with as small a P(E|-H) as possible (Howson, 2000, p. 182).

So, according to Bayesian confirmation theory, good evidence is evidence that both supports the hypothesis and counts against plausible rival hypotheses.

3.3.2. Popper's Severe Tests: Ruling out Rival Hypothese

The Popperian view of how hypotheses are corroborated is very different from Bayesian confirmation. Nonetheless, we will see that Popper shares an implicit commitment to the view that strong supportive evidence is incompatible with *and* counts against plausible rival hypotheses.

To Popper, hypotheses or theories cannot be confirmed (very roughly because of the problem of induction), but they can be *falsified* by modus tollens[38]. If a hypothesis 'survives' an attempt to be falsified, then it has been, in Popper's words, 'corroborated' (Popper 1968, section 6).

To Popper, the more *testable* (=falsifiable) the hypothesis, the higher its degree of corroboration: "*Thus confirmability (or attestability or corroborability) must increase with testability*" (Popper 1969, 11.2, emphasis original). The degree of *testability*, in turn, depends on the *severity of the test* to which the theory or hypothesis is put. A severe test is one that involves 'risky predictions' – predictions that, given our current state of background knowledge, would not bear out. Popper uses the examples of Newton and Einstein to illustrate his point:

> Newton's theory, for example, predicted deviations from Kepler's laws (due to the interactions of the planets) which had not been observed at the time. It exposed itself thereby to attempted empirical refutations whose failure meant the success of the theory. Einstein's theory was tested in a similar way (Popper 1969, 11.2).

It is clear that the *corroboration* of Newton's theory involved evidence that was incompatible with the most widely accepted rival at the time (Kepler's) *and* that the corroborating evidence for Newton's theory falsified the main rival at the time, namely Kepler's.

To use an example from Popper of tests that are not severe, consider the case of Adlerian psychoanalysis:

---

[38] (1) If the hypothesis is true, then we will observe such and such evidence (if *h* then *e*).

(2) We do not observe the evidence (not-*e*)

(3) Therefore, the hypothesis is false (not-*h*)

> As for Adler, I was much impressed by a personal experience. Once, in 1919, I reported to him a case which to me did not seem particularly Adlerian, but which he found no difficulty in analysing interms of his theory or inferiority feelings, although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. 'Because of my thousandfold experience,' he replied; whereupon I could not help saying: 'And with this new case, I suppose, your experience has become thousand-and-one-fold.' (Popper 1969, 1.1).

Adler did not submit his theory to a severe test because the evidence he cited in its favour was perfectly compatible with (at least to Popper) were non-Adlerian theories.

Popper introduces some formal vocabulary to describe the severity of a test. Where e is the evidence or test, h is the hypothesis or theory, while b is background knowledge and initial conditions, the severity of the test, S, described as a function of the evidence, the hypothesis, and the background assumptions, is[39]:

$$S(e, h, b) = p(e|h\&b) - p(e|b)$$

For simplicity I will consider cases where the hypothesis entails the evidence. In this case $p(e|h\&b) = 1$ and the measure of the severity of the test becomes:

$$S(e, h, b) = 1 - p(e|b)$$

That is, the severity of the test is equivalent to the magnitude of the difference between 1 and the probability of the evidence arising given background knowledge alone.

---

[39] Popper uses $p(x,y)$ to represent the conditional probability of x given y, and $p(xy)$ to represent the joint probability of x and y. I use the more conventional terminology. He also uses other different measures of severity, the other one being $S(e, h, b) = p(e|h, b) / p(e|b)$, and $S(e, h, b) = [p(e|h\&b) - p(e|b)] / [p(e|h\&b) + p(e|b)]$ (Popper 1969, addendum 2). It is interesting that the first of these is remarkably similar to the likelihood ratio, a Bayesian measure of the strength of the evidence. Omitting background knowledge for simplicity, the Bayes factor in favour of h (and against $\neg h$) is:

$$\frac{p(e \mid h)}{p(e \mid \neg h)}$$

With the assumption that the evidence is entailed by the hypothesis the likelihood ratio becomes

$$\frac{1}{p(e \mid \neg h)}$$

gives us a further measure of the severity of the test. The higher the likelihood ratio is above 1, the stronger the evidence.

Non-severe tests – that is, ones whose outcomes are already predicted by 'background knowledge' (and remember that this will include other theories that already seem to have evidence in their favour) – have severity 0. They are entailed by h, but they already were entailed by 'background knowledge' before h arrived on the scene. For example, the probability that someone's cold symptoms disappear within 2 weeks after taking vitamin C given background knowledge alone is almost 1. Thus, the severity of a test that measures whether people's cold symptoms disappear within 2 weeks of taking vitamin C is close to 0. Severe tests, on the other hand, are tests whose outcomes would be very low based on background knowledge alone.

The problem is, of course, determining what the probability of the evidence arising due to background knowledge alone. Besides old evidence and initial conditions, Popper intends background knowledge to mean plausible rival theories. When discussing his notion of the degree of confirmation, he describes 'background knowledge' in the following way: "Here *z* should be taken as the general 'background knowledge' (the old evidence, and old and new initial conditions) including, if we wish, accepted theories" (Popper 1969, 11.6). Or, consider the following passage:

> A theory is tested not only merely by applying it, or by trying it out, but by applying it to very special cases – cases which it yields results different from those we should have expected without that theory, or in the light of other theories [plausible rivals]. In other words we try to select for our tests those crucial cases in which we should expect the theory to fail if it is not true (Popper 1969, 3.5).

Of course it does not follow from the fact that (i) good evidence is not entailed by other hypotheses (and hence furnishes us with a severe test), that (ii) good evidence *counts against* rival hypotheses which is what 'scientific common sense' requires. At first glance, it seems that Popper was only committed to the first claim. However, close examination reveals that the very best attempts to falsify a new theory were attempts that also falsified the rival theories. Popper's description of how Newton's theory replaced Kepler's (see above), and of how the Eddington observations led to the adoption of Einstein's theory illustrates this point:

> Take one typical instance – Einstein's prediction, just then confirmed by the findings of Eddington's expedition. ... Now the impressive thing about this case is the *risk* involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted. The theory is *incompatible with certain possible results of observation* – in fact with results which everybody before Einstein would have expected (Popper 1969, 1.1.i).

The reason Eddington's observations counted so strongly in favour of Einstein's theory was not merely the fact that Newton's laws would not have predicted the observations, but that Newton's laws predicted very different observations. Hence the observations served both to support Einstein's theory and falsify Newton's laws.

In fact, there is a sense in which the fact that rival hypotheses *do not* entail the evidence (and that the evidence thereby provides us with a severe test), implies that this evidence *counts against* the rival hypotheses. In order to count as a rival hypothesis as far as a particular test is concerned, the rival hypothesis must entail that the prediction of the new theory will *not* obtain, which is logically equivalent to the claim that if the prediction bears out, that the rival hypothesis has been falsified. Expressed this way it is clear that the degree of corroboration is directly linked to whether an observation – some piece of evidence – counts against plausible rivals.

To be sure, the problem that there are, at least in principle, an infinite number of hypotheses to be ruled out, persists. But once again the aim of this work is not to solve the more general problem of underdetermination, but rather to illustrate how the Popperian account of scientific method, like the Bayesian account, appears to concur in the belief that good evidence rules out plausible rivals. To be sure, many would dispute that there were important similarities between Popperian and Bayesian accounts of evidential support (Schlesinger 1995; Gillies 1986, 1990). However, the above exposition makes it quite clear that with respect to the importance of ruling out potential rival hypotheses, the two accounts do indeed seem to converge.

3.3.3. Mill's Methods: ruling out rival hypotheses again

Mill's Methods rely overtly on the view that good evidence rules out rivals:

> The Method of Agreement stands on the ground that whatever can be eliminated, is not connected with the phenomenon by any law. The Method of Difference has for its foundation, that whatever cannot be eliminated, is connected with the phenomenon by a law (Mill 1843[1973], p. 432).

Elaborating on Mill, Mackie makes the point more explicitly: "all [of Mill's] methods work by eliminating rival candidates for the role of cause" (Mackie 1974`, p 297).

Mill's methods include the Method of Difference (which includes the Method of Residues and the Method of Concommitant Variations), the Joint Method of Agreement and Difference, and the Method of Agreement. For present purposes it will suffice to outline the Method of Difference.

Mill uses capital letters, such as "A", "B", and "C" to represent antecedents, or causes, and lower case italicized letters such as "*a*", "*b*", "*c*", to represent consequents, or effects. Mill defines the Method of Difference as follows:

> *Method of Difference*: If an instance [A] in which the phenomenon under investigation [X] occurs, and an instance[B] in which it does not occur, have every circumstance in common save one[Y], that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon [X] (Mill, III.viii.2).

Schematically,

(1) ABC → abc

(2) BC → bc

(3) Therefore, A → a

For example, "When a man is shot through the heart, it is by this method we know that it was the gunshot which killed him: for he was in the fullness of life immediately before, all circumstances being the same, except the wound" (Mill, III.viii.2). Or, to continue with the example of vitamin C, if we give vitamin C to one group of people, and withhold it from another group that is identical to the first in every respect, and those in the first group get fewer colds than those in the second group, we can conclude that the cause was taking vitamin C.

Yet again, however, the problem that there are in principle, an infinite number of rival hypotheses, remains, especially in medicine. In response to this situation, Mill can be interpreted as claiming that the best we can do is rule out *plausible* rivals. "It is true that this similarity of circumstances needs not extend to such as are already known to be immaterial to the result" (Mill 1843[1973], p. 432). Implicit in Mill's reaction is that the problem of Donald Rumsfeld's 'unknown unknowns' – the ones that might for all we know play a role but we can't judge because we don't (or don't yet) even suspect them because they have not been articulated. Concerning these, Mill was happy to admit that (at least in medicine), science was fallible. In medical experiments with human subjects, even the experimental use of the Method of Difference will not provide certain results.

> Suppose that mercury does tend to cure the disease, so many other causes, both natural and artificial, also tend to cure it, that there are sure to be abundant instances of recovery in which mercury has not been administered: unless, indeed, the practice be to administer it in all cases; on which supposition it will equally be found in the cases of failure. ... Neither, therefore, will the instances of recovery agree in the administration of

> mercury, nor will the instances of failure agree in its non-administration. It is much if, by multiplied and accurate records from hospitals and the like, we can collect that there are rather more recoveries and rather fewer failures when mercury is administered than when it is not; a result of very secondary value even as a guide to practice, and almost worthless as a contribution to the theory of the subject" (Mill 1843[1973], p. 503).

In short, the Method of Difference is difficult to use in medical research because of the problem in finding two instances with only a single difference. "In phenomena so complicated it is questionable if the two cases, similar in all respects but one, ever occurred; and were they to occur, we could not possibly know that they were so exactly similar" (Mill 1843[1973], p. 506). Mill's warning about the use of his Methods in medical research can be viewed as a warning that the potential rival hypotheses are difficult to rule out in medical cases. This view is echoed by Mackie:

> Suppose that a new drug is being tested. It is administered to some subject, and some change (good or bad) is notice in the subject soon afterwards. There is a prima facie case for supposing that the administration of this drug can cause – that is, is an inus condition of – that change. But why, if the method of difference is a demonstrative method, is it only a prima facie case? Simply because the experimenter cannot be sure that the requirements for that method's observation have been met: some other relevant change may have occurred at about the same time. But if the experiment is repeated and we keep on getting the same result, it becomes less and less likely that on each occasion when the drug was administered, some one other, unnoticed but relevant, change also occurred (Mackie 1974`, p. 316).

To sum up the discussion of Mill, his Methods are explicitly eliminative. In response to the charge that there are in principle an infinite number of rival hypotheses, Mill seems to admit that the implausible rivals can be ignored, but that we must admit the strict fallibility of our methods.

3.3.4. Further support for 'scientific common sense'

Others besides Bayesian, Popperian, and Millian accounts of scientific method affirm the importance of ruling out rival hypotheses. The idea behind Bacon's (and Newton's and Lakatos', etc.) crucial experiments is that the main rival hypotheses are ruled out by a crucial experiment. Cartwright, when read carefully, can also be found to support the idea that good evidence rules out plausible rivals. When discussing probabilistic causality, she states:

> If you see a probabilistic dependence and are inclined to infer a causal connection from it, think hard about all the other possible reasons that that dependence might occur and eliminate them one by one. And when you are all done, remember – your conclusion is no more certain than your

> confidence that you really have eliminated all the possible alternatives (Cartwright 2007, p.79).

Sir Austen Bradford Hill, arguably the Godfather of statistical methods in medicine, states: "The interpretation of statistical data turns, it should be seen, not so much on technical methods of analysis as on the application of *common sense* to figures and on elementary rules of logic" (Hill and Hill 1991`, p. 277`, italics added). Then, before concluding that one variable was the cause of another, he demands that we ask: "*is there any other way of explaining the set of facts before us? Is there any other answer which is more likely than cause and effect?*" (Hill and Hill 1991, p. 277, italics original).

Then, as we shall see below, the motivation for the belief that RCTs provide the best evidence is that they rule out plausible rival hypotheses. In short, the view that strong evidence both supports the experimental hypothesis and counts against plausible rivals seems to be very widely supported.

3.4. Towards a New Standard of Evidential Support

The accounts of scientific method I have reviewed differ in their detailed descriptions of how evidence is gathered. Nevertheless it is striking that, starting from these disparate viewpoints, they all rate strength of evidence by how well it rules out rival hypotheses. This convergence has a powerful cumulative effect that should increase our conviction that 'scientific common sense' is indeed the standard by which we should measure both types of evidence (RCTs or observational studies) as well as the strength provided by a particular piece of evidence. In addition, the methods listed above acknowledge to various degrees that science is fallible.

If we agree that 'scientific common sense' rather than hard-and-fast rules or hierarchies is the best guide to judging the strength of evidence, then several benefits follow. First, we have a rationale to evaluate and re-interpret any hierarchy of evidence in cases where a strict interpretation leads to the counterintuitive results mentioned in the previous chapter. Second, perhaps the hierarchy itself can be given a more solidly justified, if altered, form. As we saw in the last chapter, established EBM-hierarchies give special weight to double blinding in a way that automatically implies that certain types of proposed therapy can never be supported by first-rate evidence. Similarly, the view that placebo controlled RCTs are methodologically superior to active controlled RCTs is widely held. How do these issues look when we examine it from the more

commonsense perspective articulated and defended in this chapter? This is the issue that will be examined in later chapters.

My next task, however, will be to examine the concept of the placebo more critically. As will soon become evident, the terms 'placebo' and 'placebo controls' have not been adequately defined. Without such definitions, it will be difficult to compare the relative merits of 'placebo' and 'active' controls. I will therefore dedicate the next two chapters to providing adequate conceptualizations of 'placebos' and 'placebo controls'.

4. Chapter Four. Placebos as Treatments Without Characteristic Features

***Dr. Hibbert**: Why the only cure is bed rest. Anything I gave you would only be a placebo.*
***Woman in mob**: Where do we get these placebos?*
- The Simpsons

Much of the literature about the placebo effect is, in effect, an effort to debunk, confuse, or minimize it ... . Efforts to try to actually move forward our understanding of this fundamental human phenomenon are very rare
- (Moerman and Jonas 2002)

4.1. Introduction

Suppose we wish to evaluate whether placebo controlled trials are superior to 'active' controlled trials. To accomplish this, we need to know what a placebo is. Although the term 'placebo' is a common, non-technical term, and placebo controls are used in many clinical trials, whether or not we have an adequate general definition of the notion of a placebo is still a matter of debate. Asbjørn Hróbjartsson and Peter Gøtzsche were commissioned by the Cochrane Collaboration to write 7 articles on the placebo in 1994. Their first article concluded that the placebo concept is illogical: "the placebo concept ... cannot be defined in a logically consistent way and leads to contradictions" (Gøtzsche 1994). Rather than search 'blind' alleys in search of a definition of the placebo, they suggest we adopt the 'methodological' rule of taking placebos to be whatever treatments are described as placebo controls in clinical trials (Hróbjartsson 2002). Using this strategy, they undertook a meta-analysis of those trials that contained test treatment, 'placebo', *and* no-treatment groups, and found no significant placebo effect (calculated as the difference between placebo and no-treatment (Hróbjartsson and Gøtzsche 2001, 2004a). However, Hróbjartsson and Gøtzsche's strategy puts the cart before the horse. If we make the reasonable assumption that the architects of placebo controls rely on some definition of the placebo (at least implicitly), and the concept of the placebo is confused, then this "metholodogical" strategy may have lead to erroneous estimates of placebo effects[40].

---

[40]Indeed the estimates of the placebo effect gained by using this method (that overall there is no placebo effect) has been the subject of much controversy (see chapter 5 for a more complete discussion and references).

Indeed Hróbjartsson and Gøtzsche may have overlooked a good candidate for a definition of the placebo. In an influential series of papers (Grünbaum 1986, 1981), Grünbaum argued that the common definitions of placebos as inactive or nonspecific treatments are flawed. And he provided a different more sophisticated definition, which, he argued, did not suffer from those flaws. Although Hróbjartsson and Gøtzsche cite Grünbaum (Hróbjartsson 2002), admitting that it is "by far the best proposal so far", they reject his definition without much argument – devoting only 2 sentences to it – claiming that Grünbaum's definition still fails to be 'satisfying'(Hróbjartsson 2002, p. 432).

In fact, Grünbaum's paper generated a discussion of the placebo that is ongoing. In this chapter I will defend a revised version of Grünbaum's scheme that resists the important criticisms (Greenwood 1997; Hróbjartsson 2002; Waring 2003).

I begin this chapter by outlining Grünbaum's arguments that other widely accepted definitions of placebos as non-specific or inactive treatments (Shapiro and Morris 1978) are fundamentally flawed. I argue that these flawed definitions are what motivated Hróbjartsson and Gøtzsche to conclude that the placebo is illogical. In the following section, I outline Grünbaum's conceptual scheme, and endorse it subject to two minor modifications. I then consider three unsuccessful criticisms (from Greenwood and Waring) of Grünbaum's scheme. I then consider, and react to, a more serious objection made by (Hróbjartsson 2002). I conclude that the modified conceptual scheme provides an adequate definition of placebos that is interesting in its own right, and (more relevantly for the purposes of this thesis) provides the backdrop against which placebo controls can be defined.

4.2. The Powerful Placebo: in Search of a Definition

The term 'placebo' is the Latin for 'I shall please'. It is commonly used rather loosely to describe treatments that can perform the function of inducing the expectation or belief that one is being treated but that are missing some key ingredient that would make it a nonplacebo. Sugar pills, for example, prescribed in such a way that the patient believes that they are powerful painkillers, are placebos. This loose characterization of placebos has proved difficult to translate into a logically clear and watertight definition.

The Oxford English Dictionary defines the placebo as a "drug, medicine, therapy, etc., prescribed more for the psychological benefit to the patient of being given treatment than for any direct physiological effect" (OED 1989). This definition is based, as so much confusion in this area is (of course unwittingly), on a Cartesian dualist view, which makes the placebo effect appear much more mysterious than it needs to. Indeed placebo effects are perfectly compatible with the current scientific view that *there is only* physiology - psychology is simply a branch of physiology. *Moreover* this definition makes it impossible for any treatment or intervention for psychological disorders to count as anything other than a placebo. Antidepressant, and antipsychotic drugs are clearly therapies "prescribed .. for [their presumed] psychological benefit to the patient", but surely are not automatically to be characterized as placebos simply on that account. It may turn out under critical empirical investigation that this is what they indeed are – that is that patients would do just as well if prescribed a sugar pill they thought was an antidepressant or antipsychotic, but this is an empirical issue not one to be decided in advance of empirical examination by definitional fiat.

The OED goes on to define the "placebo *effect*" as "the beneficial (or occasionally adverse) effect on health produced by a placebo that cannot be attributed to the properties of the placebo" (OED). But if the effect cannot be attributed to the properties of the placebo, then to what *can* the effect be attributed? Presumably the OED has in mind implicitly a restriction of the placebo's 'properties' to its 'purely pharmacological' ones. But if, as would surely seem sensible, we characterize a placebo in terms of the full treatment regime, then one of such a treatment's properties might well be to induce an expectation of a good outcome in the patient being treated.

In a similar, misleading vein, some authors equate the placebo with an "inactive" treatment[41]. This definition is unsatisfactory because if placebos were inactive, then there would be no cases where placebo treatments outperformed 'no treatment' groups, and even sceptics about the placebo effect admit that, at least sometimes, placebo controls do outperform 'no treatment' (Hróbjartsson and Gøtzsche 2001, 2004a). Moreover, if they were always inactive (and known to be so by

[41] For example, the first sentence in one study is: " One systematic review found that pelvic floor muscle exercises increased cure or improvement rates compared with no treatment, placebo, or inactive treatments" (Onwude 2005).

definition!) what would be the point of placebo controlled trials? We could control just as well by invariably using a no treatment ('natural history') group.

Recent work on the placebo concept begins essentially with Arthur Shapiro, who spent decades studying the placebo and came up with a seemingly more defensible set of definitions – quoted here from his joint work with Morris:

> A *placebo* is defined as any therapy or component of therapy that is deliberately used for its non-specific, psychological, or psychophysiological effect, or that is used for its presumed specific effect, but is [in fact] without specific activity for the condition being treated.
> A *placebo*, when used as a control in experimental studies, is defined as a substance or procedure that is without specific activity for the condition being evaluated....
>
> The *placebo effect* is defined as the psychological or psychophysiological effect produced by placebos (Shapiro and Morris 1978, p. 371).

Shapiro and Morris's definitions are, however, subject to serious objections that Grünbaum is careful to point out. First of all they share the second defect of the OED definition: defining the placebo as a treatment used for its psychological or psychophysiological effects seems to imply that any therapy designed to treat psychological disorders automatically (and without empirical investigation) counts as a placebo. Moreover, the idea that something counts as a placebo if it achieves any effect through '*non-specific* activity' is even more problematic.

Shapiro and Morris characterize a specific activity as one where "the therapeutic influence [is] attributable solely to the contents of processes of the therapies rendered. The criterion for specific activity (and therefore the placebo effect) should be based on scientifically controlled studies" (Shapiro and Morris 1978, p. 372). In this view, then, non-specific activity would clearly seem to be the therapeutic influence *not* attributable to the 'contents of processes of the therapies rendered'. Bringing in processes alongside contents seems to obscure things even more – the process of giving what is represented as a beneficial treatment is after all supposed to be the main source of the placebo effect.

> Mindful of the potential side effects of tranquilizers and analgesics, the doctor decides to employ a little benign deceit and gives *B* a few lactose pills, without disabusing *B* of his or her evident belief that he or she is receiving a physician's sample of analgesics. Posit that shortly after *B* takes the first of these sugar pills, the headache disappears altogether. Assume further that *B*'s headache would not have disappeared just then from mere internal causes. ... Thus *B* assumedly received the same headache relief from the mere sugar pill as he or she would have received if a

> pharmacologically *non*inert drug had been slipped into his food without his knowledge.
>
> Clearly, in some situations, the therapeutic effect of the sugar pill placebo on the headache can have attributes fully as sharply defined or 'specific' as the effect that would have been produced by a so-called 'active' drug .... Moreover, this placebogenic effect can be just as precisely described or known as the nonplaceboegenic effect of aspirin. In either case, the effect is complete headache relief" (Grünbaum 1986, p. 31).

Indeed there are actual examples similar to the one Grünbaum describes, and the mechanism through which placebos could 'kill' pain have been described. In brief, there is evidence that placebos may increase the level of endogenous opiods (Benedetti, Rainero, and Pollo 2003).

A second possible sense of 'specific' refers to the following type of situation:

> the remedial effectiveness of *t* is specific to a quite small number of disorders, to the exclusion of a far more multitudinous set of nosologically different afflictions and of their respective pathognomonic symptoms" (Grünbaum 1986, p. 31).

That is, a non-placebo on this account would count as 'specific' because it has remedial effects for a limited number of disorders. For example, fluoxetine is supposed to have remedial effects only for depression and related disorders. Sugar pills ('non-specific' treatments on this view), on the other hand, can – allegedly - function as placebos for most ailments. But this sense of the term is also misleading. If Hróbjartsson and Gøtzsche were right the placebo effect would in fact be specific just to pain (admittedly to pain from various sources, but then how specific must specific be?).

A variation of this definition of placebos as non-specific is the idea that placebos are the common factors of treatments. First, the proponents of this view fail to state what the common factors are (Grünbaum 1986, p. 33). More importantly, the view falls prey to similar objections to the second interpretation of the term 'specific'.

Secondly, the term 'specific' is sometimes used to denote 'well-defined' as "when Miller (1980) writes that 'placebo effects can be quite specific'" (Grünbaum 1986, p. 32). This third sense in which the term 'specific' is used seems to contradict the first two senses. In the first two, non-placebos but not placebos are specific, while in the third, placebos *are* specific. Even (Miller and Chilton 1980)whom Grünbaum uses as an example of someone who understands 'specific' in this third sense, trips himself up with various contradictory definitions: "the illustrations he [Miller] goes on to give show that here 'specific' has the force of 'quantitatively precise'. But in the very next paragraph, he uses the term 'specific' as a synonym for 'nonplacebo' when reporting that 'it is only

in the past 80 years that physicians have been able to use an appreciable number of treatments with specific therapeutic effects'" (Grünbaum 1986, p. 32).

To sum up this section, Grünbaum claims that "this standard technical vocabulary [of placebos as nonspecific or inactive treatments]... generates confusion by being misleading or obfuscating, and indeed cries out for conceptual clarification" (Grünbaum 1986, p. 27). In particular, "the generic distinction between placebos and nonplacebos has nothing whatever to do with the contrast between nonspecificity and specificity" (Grünbaum 1986, p. 33). And on these issues he seems to be correct.

4.3. Grünbaum's scheme: the necessity of *relativizing* the definition of placebo to a *therapeutic theory*

Grünbaum's main point is that trying to find a characterization of what counts as a placebo *simpliciter* is bound to fail, because whether some intervention counts as the administration of a placebo does not just depend on what that intervention is, but on the target disorder D and on the therapeutic theory, ψ, underpinning whatever is being taken to count as a *non*-placebo for D. Here is why we need to relativize to target disorder D:

> Ironically, none other than the much-maligned proverbial sugar pill furnishes a *reductio ad absurdum* of the notion that a medication can be generically a placebo *simpliciter*, without relativization to a target disorder. For even a lay person knows that the glucose in the sugar pill is anything but a generic placebo if given to a victim of diabetes who is in a state of insulin shock, or to someone suffering from hypoglycaemia. (Grünbaum 1986, p. 35).

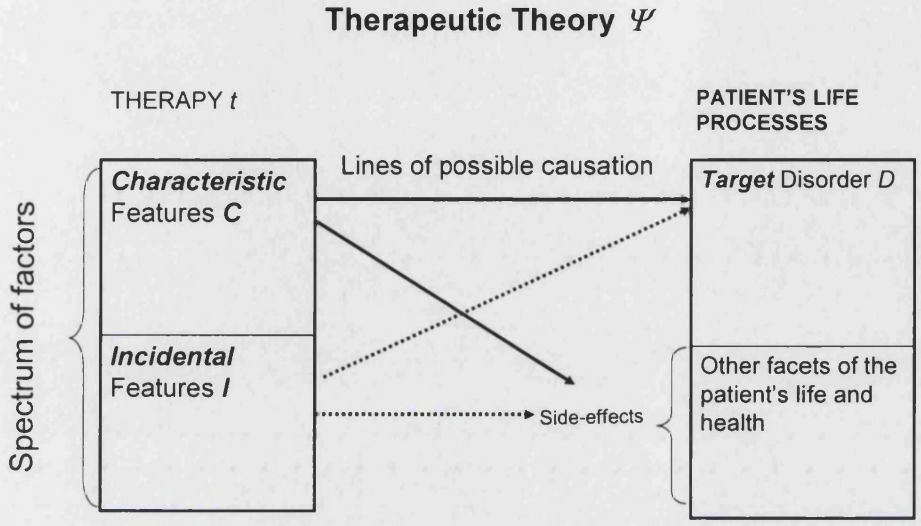
Grünbaum also holds – rather less obviously – that we also cannot arrive at an adequate account of a placebo unless we note a dependence not just on the target disorder but on the therapeutic theory. The main concern here is the recognition that a particular treatment may be a non-placebo overall and yet involve aspects or features that are plausibly characterized as placebic. Grünbaum records that, for example, there is some evidence that chemotherapy for certain kinds of cancer is enhanced in its positive effects if administered by an enthusiastic physician. The theory of the effects of chemotherapy on tumours would then dictate which characteristics of the overall treatment are '*characteristic*' - that is, those 'treatment factors that a given theory ψ... picks out as the defining characteristics of a given type of therapy t – and any other factors that are as a matter of fact involved in the treatment are called 'incidental factors'. Similarly in the case of using Prozac as a treatment for depression, the characteristic factors, according to accepted theories, are the pharmacological effects of

the fluoxetine on serotonin levels in the brain, while incidental factors would include the water used to help swallow the pill and the details of the patient/clinician interaction.

Often in 'somatic medicine' as Grünbaum calls it (though this again tends to encourage unfortunate dualist tendencies), there is so little controversy over the therapeutic theory presupposed that it might seem artificial to talk about a theory at all. To take an example Grünbaum cites the 'theory' that underwrites accepted treatment for gallstones will clearly make the surgical removal of the gallstones (as opposed, say, to a sham incision) as characteristic. But in the psychotherapeutic field (which initially motivated Grünbaum's concern with the placebo concept) the dependence on theory is often crucial. In that field, which aspects of a particular interaction with a patient are characteristic will clearly be theory-dependent to the extent that one and the same feature of a given interaction may be judged characteristic by one theory and incidental by another. For instance, taking a patient's history and exploring their past could well be characteristic for some forms of psychoanalysis but incidental for more 'action orientated' cognitive behaviour therapy (CBT).

Rather than specify the particular properties of the placebo *simpliciter*, which runs into the problem of classifying the sugar pill as a necessary placebo, the relevant medical or psychological therapeutic theory tells us what a placebo (or rather a 'placebic element' in some overall treatment) is in each case (Grünbaum 1986, p. 29-30). Grünbaum's scheme is best explained with the aid of a diagram:

2. Diagram 4.1: Illustration of therapeutic theory ψ, used in clarifying the definition of 'placebo' (Grünbaum 1986, p. 22)



Beginning with the left-hand box in the diagram, we see that the therapeutic theory, ψ, differentiates between characteristic (C) and incidental (*I*) features[42]. This is indicated on the left-hand box of the diagram. For example, a therapeutic theory may state that the therapy *t* is the administration of Prozac according to some given regime, the target disorder *D* being major depressive disorder (MDD). The therapeutic theory might specify that the physiological effects of the chemical fluoxetine (the name of the patented compound in Prozac) are the 'characteristic features', *C,* of this therapy. The incidental features, *I,* of the therapy might include,e.g., the water with which the Prozac pills are swallowed and, clearly potentially more significantly, the patient/doctor interaction. Or, to use Grünbaum's example, "a theory that deems the removal of gallstones to be therapeutic for certain kinds of pains and indigestion will assume that this abdominal surgery includes the administration of anaesthesia to the patient" (Grünbaum 1986, p. 22). In this case, the anaesthesia would be an incidental feature, and the removal of the gallstone would be a characteristic feature.

The right-hand box in the diagram divides the patient's life processes into two parts. First, there is the target disorder *D* at which the therapy *t* is aimed; second, there

---

[42] For some reason, Grünbaum uses *F* to designate characteristic factors and *C* to designate incidental factors. Because *C* for the former and *I* for the latter is more natural, I will adopt it.

are other facets of the patient's life. I will supplement Grünbaum's scheme and refer to these other factors collectively as *O*. Any effects the therapy has on these other processes (whether beneficial or harmful) Grünbaum calls side-effects[43].

Lastly, the four arrows in the diagram represent *possible* effects. The top horizontal arrow represents the possible effect of the characteristic factors *F* on the target disorder *D*. The arrow that runs top half of the left-hand box to the lower half of the right-hand box represent the side effects of the characteristic factors. The lower horizontal arrow represents the side effects of the incidental factors *I*, while the remaining arrow represents effects of the incidental factors on the target disorder *D*. The four arrows of possible causal influences can be positive, negative, or, in some cases 'empty' i.e. represent no effects at all. "Thus, one or more of the characteristic factors *F* may be remedial for the target disorder *D*, or the *F* factors may have no effect on *D*, or the *F* factors conceivably could make *D* even worse" (Grünbaum 1986, p. 22).

The conceptual scheme, in order to tell us what *actually* counts as a characteristic or placebic effect must be 'filled in' according to the therapeutic theory, which will specify the target disorder, and the characteristic and incidental features. "It is therefore not my explication but a given theory ψ that determines which treatment factors are to be classified as the characteristic factors in any one case" (Grünbaum 1986, p. 29). For example, "the given therapeutic theory ψ (in medicine or psychiatry) rather than my explication determines whether any factors in the physician-patient relationship are to count as only 'incidental'" (Grünbaum 1986, p. 29). This, he claims, avoids the confusions "generated when investigators want to assess the generic placebo status of a therapy *t* across rival therapeutic theories, and without regard to whether these theories use different characteristic factors to identify *t*" (Grünbaum 1986, p. 30). Grünbaum can then go on to define what a placebo is and in fact finds it best to start with the notion of a *non*-placebo:

---

[43] It could be argued that the act of taking a treatment is a consequence of any treatment. However this is not necessarily a *negative* side-effect. Some people enjoy taking their pills, and some treatments do not involve any action on the part of the patient. In one case, for example, hospitalized patients were prayed for 'retroactively', that is *after* they had left the hospital (Leibovici 2001). Whether or not such a treatment could be effective is another matter, the point is that not all treatments can be said to have the negative side effect of taking the treatment.

Nonplacebo (1): a treatment *t* is a nonplacebo "*if (and only if) one or more of the characteristic factors do have a positive therapeutic effect on the target disease D,*" (Grünbaum 1986, p. 23).

The administration of Prozac, for example, would be characterized as a nonplacebo for depression if and only if fluoxetine (the characteristic feature of treatment with Prozac according to the therapeutic theory that endorses Prozac) has some *positive* therapeutic effect for depression. That is, Prozac would be considered a nonplacebo if the top horizontal arrow represented a positive effect. Grünbaum does not, therefore, require that the effects of a nonplacebo are due *solely* to the characteristic factors, and indeed he explicitly acknowledges that incidental factors may enhance the effects of characteristic factors (Grünbaum 1986, p. 28). However it is should be noted that the characteristic features of a treatment process *t* must, on G's account, have a *positive* therapeutic effect in order for *t* to be a nonplacebo. G then proceeds on this basis to characterise the notion of a 'generic placebo':

Generic Placebo: a treatment process *t* is a placebo if none of the characteristic treatment factors *C* are remedial for *D* (Grünbaum 1986, p. 33).

Given this understanding of the term, it is clear, as Grünbaum himself goes on to emphasise, that generic placebos come in two types: intentional and inadvertent.

Grünbaum characterises the first type as:

Intentional placebo: a treatment process *t* is an *intentional placebo* if and only if it satisfies the following four conditions:

(a) ***it is a generic placebo***

(b) ***the practitioner*** **believes** ***that t is a generic placebo***. "*P* [the practitioner] believes that the factors *F* [*C*] indeed all *fail* to be remedial for *D*" (1986, p.24).

(c) ***the practitioner believes that some patients will benefit from the treatment due to one or more of its incidental features*** . "*P* also believes that – at least for a certain type of victim ... of *D* – [the treatment] is nonetheless therapeutic for *D*" (1986, p.24), and

(d) **[optional]** ***the victims of D who are treated with t believe that t has some remedial*** **characteristic** ***features***. "*P* abets, or at least acquiesces in, [the victim's] belief that *t* has remedial efficacy for *D* by virtue of some constituents that belong to the set of characteristic factors *F* [*C*]" (1986, p.24).

Grünbaum characterises the second type of placebo as:

Inadvertent placebo: a treatment process *t* is an *inadvertent* placebo if and only if it satisfies the first two of the following three conditions – the third normally holding but, strictly speaking, being optional:

> (a) ***t is a generic placebo***
>
> (b) ***the practitioner believes that some of the characteristic features are remedial for D***: "at least for a certain type of victim ... of *D* – *P* credits these very ['characteristic'] factors [*C*] with being therapeutic for *D*" (1986, p.28).
>
> (c) [like (d), above, this condition is optional] ***the patient believes that the remedial effects on D are due to some characteristic feature of the treatment t.*** "More often than not, [the patient] believes that *t* derives remedial efficacy for *D* from constituents belonging to *t*'s characteristic factors" (1986, p.28).

Given the aim of this chapter, namely to provide a basis for the correct identification of placebo controls, the distinction between intentional and inadvertent placebos is irrelevant. What is important to note are the shared features of the definitions of both intentional and inadvertent placebos. To emphasise: in order to count as a generic placebo Grünbaum *only* requires that the characteristic features of the treatment have no positive effects for the target disorder.

Placebo effect: a placebo effect is either (a) one produced by the incidental features of some treatment[44]. Or (b) placebo effects are any effects of a generic placebo[45].

This means in particular, and rather oddly, that if a treatment has an effect at all, then this counts as a placebo effect just in case the treatment is not a non-placebo; and this can happen in one of two ways (a) the effect may be attributable to the incidental factors of the treatment or (b) the characteristic features may have effects, but none is positive – that is some may be neutral and at least one negative. The reason this is odd is that, in a way that strongly contravenes ordinary usage, treatments whose characteristic features have *negative* effects on the target disorder will count as generic placebos. Yet

---

[44] "even when a treatment is a *non*placebo, effects on *D* – be they good, bad, or neutral – that are produced by *t*'s *incidental* factors count as placebo effects, precisely because these factors wrought them" (Grünbaum 1986, p. 23).

[45] "when *t* is a generic placebo whose characteristic factors have harmful or neutral effects on *D*, these effects as well count as placebo effects. Hence, if *t* is a placebo, then *all* of its effects count as placebo effects" (ibid, p. 23). (Grünbaum 1986, p. 23).

this is just what Grünbaum's scheme implies: "any treatment *t* qualifies generically as a placebo for a given target disorder *D* merely on the strength of the failure of *all* of its characteristic factors *F* to be remedial for *D*" (Grünbaum 1986, p. 33). There is no requirement that the characteristic factors not be harmful for *D*. Thus, deep scratching of the skin (which is, we assume is the only characteristic factor) would be classified as a placebo for hemophilia. Such a treatment for hemophilia is surely not a placebo, if we follow common usage of the term 'placebo'.

Grünbaum supports calling harmful treatments placebos because he claims that before the dawn of modern medicine, most treatments were merely placebos, and that some were harmful. His generalization of what characterizes a placebo is "prompted by the sobering lesson of the history of medicine that most treatments were inadvertent rather than intentional placebos, and often harmful to boot!" (Grünbaum 1986, p. 33).

However, it seems like a misuse of the term 'placebo' to lump treatments that have negative effects on *D* (and not merely negative side-effects) together with treatments whose incidental features have positive effects on the target disorder. Similar arguments apply to treatments whose characteristic features have no positive effects on the target disorder but exacerbate the patient's other life processes. For example, imagine that a drug for depression (the effects of which were the only characteristic features of the treatment according to the underlying therapeutic theory ψ) for depression did not influence the depression at all, but had the side effect of paralyzing the patient. On Grünbaum's original scheme, such a treatment is a placebo, but these treatments are surely better classified as toxic agents.

I will therefore introduce the term *toxic agents* to cover *either* treatments whose characteristic features have negative effects on the target disorder, *or* treatments whose characteristic features have no effects on the target disorder, but *do* have negative side effects.[46] Red Bull, for example, could be characterized by some therapeutic theory as a

---

[46] With three overall 'values' for effects (positive, negative, and neutral), and four types of effects (*C* on *D*, *C* on *O*, *F* on *D*, *F* on *O*), there are 3^4, or 81 possible combinations of treatment types. For example, one could suggest that we give a special name for treatments whose characteristic features have mild remedial effects on the target disorder but that have horrible side effects. Although Grunbaum may have oversimplified, there is the danger of obfuscating. I therefore refrain from making further modifications. Treatments whose characteristic features

toxic agent for severe anxiety if, and only if, caffeine and taurine (the characteristic features) have negative therapeutic effect for anxiety or if they have no effect on anxiety but, say, cause people's toes to atrophy. It could be argued that my modification is unsupportable because Grünbaum objects to the requirement that placebos cannot be treatments whose characteristic features are harmful for *D*.

> [N]ote that if one were to define a generic placebo therapy *t alternatively* as one whose characteristic factors are *without [negative] effect* on *D*, it would have the consequence that a *non*-placebo *t* would either exacerbate *D* or be remedial for it, or would have a merely neutral effect on it. But in my definitional scheme, the characteristic factors of a *non*-placebo must be positively therapeutic (Grünbaum 1986, p. 33).

But Grünbaum does not follow up with an argument in support of what after all, in the end, is a semantic *decision*. Perhaps he reasons that because he has defined nonplacebos as treatments whose characteristic factors have a positive effect on the target disorder, it follows that all other treatments must be classified as placebos. One might say that he would be correct if there were a strong reason to stick to a simple 'placebo or nonplacebo' scheme. But given the consequences of such a scheme – that we end up lumping what most would call toxic agents into the placebo category, it may be more advisable, as I have suggested, to include a category of *toxic agents*, namely treatments whose characteristic features have an overall toxic effect.

A similar clarification must be made with regards to the potentially negative effects of incidental features. Imagine that a therapeutic theory deems the delivery of a medication by an authority figure in a white coat and stethoscope to be an incidental feature. Imagine further that this incidental feature generally enhances the effects of the characteristic features on anxiety disorders. However, if certain people are terribly frightened of authority figures, then this incidental feature may have a negative effect on anxiety. It seems strange to call such an effect a placebo effect, at least for certain very anxious patients[47]. Indeed the literature distinguishes between placebo effects and nocebo effects. The term 'nocebo' is latin for 'I shall harm', and refers to the negative effects produced by placebos, or incidental factors. In some places the term 'nocebo' is

---

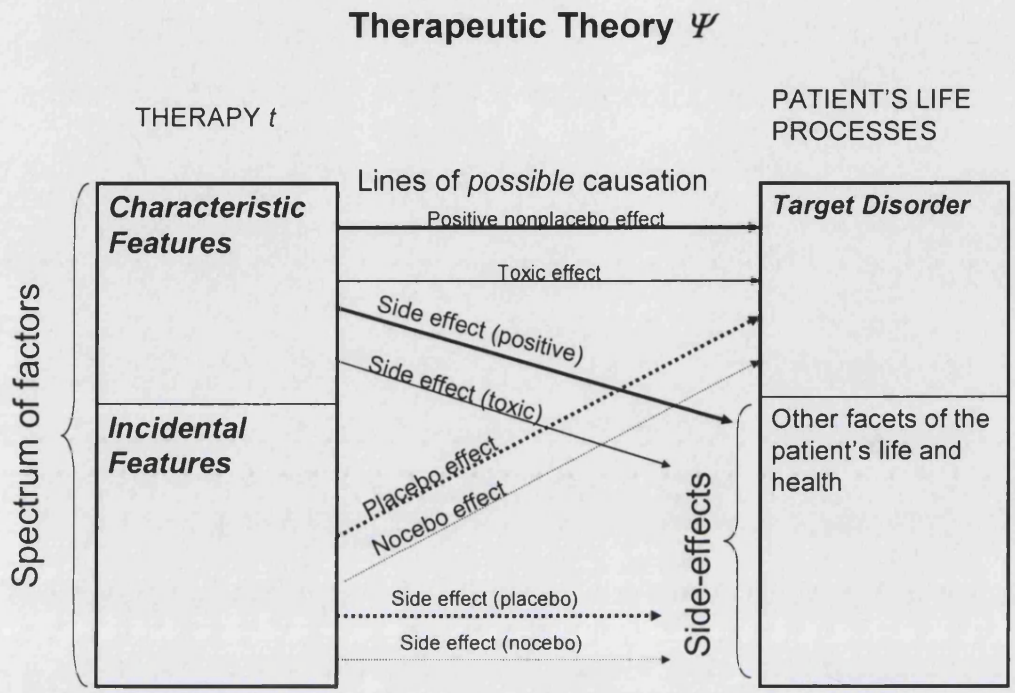
have mild remedial effects but extreme side-effects can be classified as nonplacebos with undesirable side effects.

[47] I address the problem that what counts as a placebo must be relativized to particular patients in a later section.

used to describe a treatment that produces negative *side-effects*, but there is neither general agreement nor good reason to limit the use of the term in this way. I will therefore introduce the term 'nocebo' into Grünbaum 's scheme to denote treatments whose only effects are exacerbating and produced by the incidental features.

I therefore recommend that Grünbaum's original scheme be revised to distinguish between *placebos*, where the characteristic features have no effects but that the incidental features produce a good overall effect, and *nocebos*, where there are no characteristic effects but negative incidental effects. Then, *nonplacebos* are treatments whose characteristic features have an overall positive effect on the target disorder, and *toxic agents* are nonplacebos that have an overall negative effect on the target disorder (or some other life-process). The revised scheme can be explained with the help of a figure (see below):

3. Diagram 4.2: Revised Illustration of Therapeutic theory , Used in Clarifying Definition of 'Placebo': Nonplacebo, Toxic, Placebo, and Nocebo Effects



With the revised scheme in mind, the definitions of placebos, nocebos, nonplacebos and toxic agents can be summarized as follows:

Nonplacebo (revised): a treatment *t* is a nonplacebo, relative to a target disorder D and a therapeutic theory ψ if (and only if), the characteristic factors specified by ψ have an *overall positive effect* on D.

Toxic agent: A treatment t is toxic relative to a target disorder D and therapeutic theory ψ iff the overall effect of the characteristic features of t specified by ψ is negative for D.

Placebo (revised): (a) A treatment t is a *generic* placebo relative to a target disorder D and therapeutic theory ψ iff it is neither a positively effective non-placebo nor a toxic agent but has an overall positive effect via its incidental features; (b) t is an *intentional* placebo relative to a target disorder D and therapeutic theory ψ iff (i) t is a generic placebo and (ii) t was prescribed by a clinician in the (correct) belief that any positive effects would be achieved through incidental factors (and possibly encouraged the patient to believe (falsely) that the effect would be via characteristic factors); otherwise (that is if the clinician prescribes t in the false belief that its characteristic features will have a positive effect) it is an *inadvertent placebo*. Whether it is an intentional or inadvertent placebo the victims of the disorder believe that the characteristic features of the treatment have remedial effects.

Nocebo: A treatment t is a *generic* nocebo relative to a target disorder D and therapeutic theory ψ iff it is neither a positively effective non-placebo nor a toxic agent, but has an overall negative effect via its incidental features. The treatment process *t* is an *intentional* nocebo[48] relative to a target disorder D and therapeutic theory ψ iff (i) t is a generic nocebo and (ii) t was prescribed by a clinician in the (correct) belief that negative effects would be achieved through incidental factors; otherwise (that is, if the clinician prescribes t in the false belief that its characteristic features will have a positive, rather than negative, effect) it is an *inadvertent nocebo*

Placebo effect: (unchanged, with the stipulation that placebo effects *not* be nocebo effects).

Nocebo effect: Nocebo effects are the exacerbating effects of the incidental features. It must be specified whether these effects are on the target disorder *D* or on other facets of a patient's life *O*.

There is a reason why even Grünbaum himself should accept this modification: he insists that placebo *controls* (which I consider in more detail in the next chapter) should not be toxic. He restricts placebo controls to treatments that are "generally quite harmless to those victims of *D* who have been chosen for the control group" (Grünbaum 1986). This suggests that he does not intend placebos to be treatments that are harmful. On the revised scheme any placebo is a potential candidate for a placebo control, whereas on the original scheme that is not the case. The modified scheme, I submit, is more in line with our intuitions regarding placebos and treatments whose characteristic factors are harmful.

In actual fact, most medical treatments will not fit neatly even into the more complex categories in the revised schema. Real medical treatments have both characteristic and incidental features, and both of these can have positive *and* negative effects on *both* the target disorder, or other life processes. The definitions of placebos, nonplacebos, toxic agents, and nocebos fail to reflect this. For instance, what are we to make of an 'ambiguous' treatment whose positive characteristic features on the target disorder are swamped by the negative effects (again on the target disorder) of the incidental features? We would not want to call these treatments positively effective nonplacebos, although this is what the written definitions would have us conclude. A

[48] Intentional nocebos are, to be sure, unlikely to be prescribed. I include it merely for the sake of completeness.

similar problem arises if the negative effects of a toxic agent are overwhelmed by the positive incidental effects.

Another solution would be to introduce a more complex definitional schema. Here, positively effective nonplacebos, for example, would be treatments whose characteristic features have positive effects on the target disorder *and* these positive effects were greater than the sum of the negative effects (on the target disorder) wrought by *either* characteristic or incidental features. The problem with this solution is that too many definitions would be required. For instance, what would we do with a treatment such as Thalidomide, whose characteristic features were (apparently) positively effective on the target disorder, but whose (characteristic or incidental) features had extremely negative side-effects? Surely we would not want to deem such a treatment to be a positively effective nonplacebo, yet even on the revised definitional scheme, that is what it is.

A simple solution could be to change the relativization from the target disorder to a more broadly construed target *outcome*. If the outcome of a Thalidomide trial were simply relief from morning sickness, then you might get the problems cited, but no one could sensibly think that it was. Rather, a sensible outcome would be something like 'relief of morning sickness while maintaining the health of the foetus', in which case there is no question of this being a positive nonplacebo overall.

Note that the problem of ambiguity does not plague the definitions of placebos or nocebos in the same way. A placebo is a treatment whose characteristic features are ineffective and whose incidental features have an overall positive effect. A nocebo is a treatment whose characteristic features are ineffective and whose incidental features have an overall negative effect.

Still a more robust solution to the problem of 'ambiguous' treatments is to avoid *overall* classifications and stick to the idea captured by the diagram that any treatment can have positive nonplacebo, toxic, placebo, and nocebo effects on either the target disorder or other life processes. Of course the revised version of Grünbaum's diagram represents this idea well – the arrows represent the many possible effects of a treatment. Still, it is useful to distinguish between what we call a placebo and what we call a nonplacebo. Hence it is best to avoid statements such as: 'x is a positively effective treatment' and replace it with more complete descriptions of the effects of each treatment feature, i.e. 'features of x_i (be it characteristic or incidental) have effects y_j for outcomes z_m'.

4.4. Four Failed Critiques of Grünbaum's scheme

Grünbaum's definitions have generated a still ongoing discussion. Waring (2003) argues that paradoxical drug responses present a problem for Grünbaum's scheme. Greenwood[49] (1997) argues that if positive effects can be explained pharmacologically (whether or not via any pharmacological factors deemed 'characteristic' by a therapeutic theory) then the treatment should *not* be classified as a placebo. They both claim that Grünbaum fails to emphasize that placebo effects must be pyschological. I will argue that the first objection is based on a misreading of Grünbaum's text, while the other two fall prey to the same problem with considering anything particular (such as sugar) to be necessarily placebogenic.

4.4.1. Waring's mistaken assumption that treatments are not relativized to patients in Grünbaum 's scheme

Waring uses the example of drugs that elicit 'paradoxical responses' to show how the same treatment can be classified both as a placebo and as a nonplacebo by Grünbaum. A paradoxical response is an exacerbating response on the target disorder of a drug that is normally remedial. Waring claims that drugs with paradoxical responses include antidepressants for depression and anxiolytics for anxiety. I will focus on the case of antidepressants. Waring writes:

> [C]onsider the newer generation of Selective Serotonin Reuptake Inhibitors (SSRIs). There is evidence that they might induce acutely anxious and even suicidal behavior in certain patients suffering from anxiety and depression (Waring 2003, p. 12).

Although SSRIs may be effective for *most* victims of depression they cause a worsening of depressive symptoms in others, or so Waring argues. This seems to result in Grünbaum's account being contradictory. The same treatment, t, characterized by the same therapeutic theory, ψ, is a nonplacebo for some patients and a placebo for others. "Thus, an anxiolytic that alleviates anxiety for patient a and exacerbates it in patient b would be a nonplacebo for a and a placebo for b" (Waring 2003, p. 12).

One problem is that the paradoxical (exacerbating) response of the drugs are termed placebos by Grünbaum , which seems odd: "I would regard calling such

[49] Ironically, "Grünbaum" translated into English is "Greentree" or "Greenwood".

[paradoxical] effects "placebic" as a misuse of language" (Waring 2003, p. 12). Here I concur with Waring, and have introduced the notion of toxic agents into Grünbaum's scheme (see above) to deal with it. Waring might acknowledge that a modification of the scheme deals with the oddity of calling toxic treatments 'placebos', but note that the main problem persists: the same therapeutic theory characterizes one and the same treatment as both a nonplacebo and a toxic agent.

The obvious response for Grünbaum is to require that treatments be relative not only to disorders, but also to patients. In this view, antidepressant drugs would be nonplacebos for patients in whom characteristic features have remedial effects. Then, for Grünbaum, they would be placebos for those in which the same characteristic features have negative effects. This still seems unacceptable – but only for a reason we have already dealt with. As I have already argued, we should, independently of this problem, modify Grünbaum's scheme by introducing a further category of 'toxic agent'. In that case, the 'paradoxical' drug would be categorized as a toxic agent for those patients where the characteristic features make their condition worse. Waring claims that relativizing treatments to particular patients is not an option for Grünbaum. He claims that a drug which

> alleviates anxiety for patient *a* and exacerbates it in patient *b* would be a nonplacebo for *a* and a placebo for *b*. Note the interesting consequence that some placebos would be patient relative. On Grünbaum's scheme, the ingredients of a *t* can only be reclassified as *F* or *C* under different theories, i.e., one *Y* can designate the same factors of a *t* as characteristic and another *Y* can designate them as incidental (Waring 2003, p. 12).

But it is unlikely that Grünbaum or any sensible practitioner would fail to recognize that treatments need to be relativized to patients. Most, if not all treatments, ranging from aspirin to chemotherapy to antibiotics do not work universally. They work for most people, but not for everyone. That is, whether they count as nonplacebos depends on the person. To use a more dramatic example, swimming might be a wonderful treatment for obesity or rehabilitation, or general well-being *but only for those patients who know how to swim*. Swimming could lead to death by drowning – a clear exacerbation of well-being – for non-swimmers. By necessity, the therapeutic theory must specify, in addition to which factors are incidental, which patients for which the treatment is a nonplacebo.

In fact, a careful reading of Grünbaum indicates that this is the interpretation he intended. For instance, when describing intentional placebos he makes explicit reference to particular 'victims': "A treatment process *t* ... will be said to be an 'intentional' placebo with respect to a target disorder *D*, suffered by a victim *V* and treated by a dispensing practitioner *P*" (1986, p. 24). Or later, when referring to both types of placebo (i.e. of intentional and inadvertent), he states: "Both explications are *relativized to disease victims of a specifiable sort*, as well as to therapists (practitioners) of certain kinds" (Grünbaum 1986, p.35, emphasis added). In short, Grünbaum , like any sensible practitioner, knows that treatments need to be adapted to patients. A careful reading of Grünbaum's text reveals explicit statements that therapeutic theories need to be relativized to particular patients. Once this is acknowledged, the problem highlighted by paradoxical drug responses is exposed as illusory. However, in fairness to Waring it should be allowed that when spelling out his definitions (which he does often and in great logical detail) Grünbaum never actually makes the relativization to patients explicit in his scheme and one only finds it implicit in various bits of commentary.

4.4.2. Greenwood's misidentification of pharmacologically active with nonplacebo

Greenwood argues that Grünbaum's concept of the placebo as relativized to the target disorders and therapeutic theory has absurd consequences. He states:

> In what follows I try to explain why, by documenting the unhappy consequences of Grünbaum's unrestricted *negative* definition of "placebo effects" and "generic placebo" in terms of factors designated as "incidental" according to a therapeutic theory T of treatment t for disorder D (Greenwood 1997, p.499).

In particular, if a factor in t is declared 'incidental' by ψ but is pharmacological rather than psychological while none of the factors of *t* declared characteristic by ψ has any effect, than t counts as a placebo on Grünbaum's scheme but, says Greenwood, does not do so intuitively:

> Consider the hypothetical case of a drug treatment t for disorder D. According to therapeutic theory T of drug treatment t for disorder D, the pharmacological components a, b, and c are "characteristic" or "active"

> components [C]; the pharmacological components d and e are "incidental" or "inert" components [I]. Say it turned out to be the case that components a, b and c are not remedial for D, but that component e alone is responsible for the total remedial effect. In this case, where the effect is produced by pharmacological component e alone, we would have an instance of a placebo effect, according to Grünbaum's definition *even though no part of the effect is produced by psychological factors such as therapist/doctor commitment or client/patient expectancy.* I think that to call such a *pharmacologically produced* effect a "placebo effect" is a misuse of language. Any account that has such as consequence is off to a very bad start (Greenwood 1997, p. 500)

Greenwood challenges Grünbaum's notion that placebos and nonplacebos must be relativized to a therapeutic theory ψ by maintaining that we can specify what a nonplacebo is independently of any ψ: in particular if t achieves its effect through a component that is pharmacological then it counts as a (generic) non-placebo, no matter how the prevailing ψ categorises that component.

Although Greenwood does not provide an example to illustrate the apparently unhappy consequences of Grünbam's scheme, he surely has in mind a case such as the following. Imagine some treatment for bacterial pneumonia had the following treatment factors:

a: anything with no remedial effects on the target disorder

b: anything with no remedial effects on the target disorder

c: anything with no remedial effects on the target disorder

d: patient/doctor expectancy

e: antibiotics.

Imagine further that the therapeutic theory classified *d* and *e* as incidental while *a*, *b*, and *c* were classified as characteristic. This illustrates the apparent problem with Grünbaum's scheme: it does, in fact, seem to be a misuse of language to call the imaginary treatment a placebo for bacterial pneumonia. The simple solution to the problem would be to disallow positively therapeutic, pharmacologically produced agents to be classified as incidental.

However close examination reveals that the alleged problem with Grünbaum's scheme may not be serious, while the problem with Greenwood's solution is serious. I will first focus on the problems with Greenwoods intended counter-example.

To begin, the imaginary therapy used in the imaginary example must be underwritten by a therapeutic theory. Given the current state of medical knowledge, it

would be a strange therapeutic theory indeed that characterized antibiotics as incidental for the treatment of bacterial pneumonia. In fact, Greenwood's example only reduces Grünbaum's theory to the absurd if we allow the therapeutic theory to be completely arbitrary. Although Grünbaum does not place strictures on what counts as a justified therapeutic theory (I address this issue below), it is reasonable to require current medical knowledge, as well as common sense, to play some role. Classifying antibiotics as incidental for the treatment of bacterial pneumonia surely flies in the face of current knowledge as well as scientific common sense.

Of course, antibiotic is a heavily theory-laden term, so it is unlikely that a therapeutic theory would not pick it out as characteristic for a known bacterial disorder. More realistically, 'substance X', a more mysterious substance, could be responsible for any therapeutic benefit, but the therapeutic theory could fail to pick it out as characteristic. This could lead to a therapeutic theory mistakenly classifying some pharmacologically active substance as incidental when, upon closer examination or with more information, it would be classified as characteristic. I address this problem with potentially flawed therapeutic theories in a later section.

The second problem is that antibiotics, just like most imaginable substances including sugar, have placebo effects for some disorders and nonplacebo effects for other disorders. Even powerful antibiotics are placebogenic for the viral cold, for example, while they may have strong therapeutic effects on bacterial pneumonia. Antibiotics, clearly pharmacological substances, may not be responsible for any therapeutic benefit for the viral cold. Here we have an instance of a pharmacologically produced, yet clearly placebogenic factor, which shows that Greenwood's claim that pharmacologically produced, yet positively effective substances are not necessarily characteristic.

Greenwood could reply that, in addition to being pharmacologically produced, the factor cannot produce its effect "by psychological factors such as therapist/doctor commitment or client/patient expectancy". If we ignore the possibility that this is an unacceptable, *ad hoc* addendum to his intended counterexample to Grünbaum, it neutralizes the example of antibiotics for the viral cold as a response to Greenwood.

This brings us to the third, and biggest problem for Greenwood's claim, namely that it does not, at least in general, count against Grünbaum's scheme. To see why, imagine that a therapeutic theory ψ characterizes the treatment factors *a*, *b*, *c*, *d*, *e* of treatment *t* called 'Superwater' for sore throat in the manner Greenwood suggests: *a*, *b*,

c are characteristic, while *d*, *e* are incidental, and *e* is the only factor responsible for the remedial effects of *t*, moreover *e* is pharmacological. Now imagine that the target disorder is sore throats, and the therapeutic theory assigns the following instantiation to the various treatment factors:

a: completely ineffective pharmacological compound P_1

b: completely ineffective pharmacological compound P_2

c: completely ineffective pharmacological compound P_3

d: red food colouring

e: pharmacological compound P_4, a new type of pharmacologically produced purified water whose only purpose is to wash down the pill. Tap water would have the same benefit, and the only reason this particular pharmaceutical company produces P_4 is that they are trying to break into the distilled water market.

The factors P_1, P_2, and P_3 were experimental agents that were suspected to cure sore throats but in fact had no effects whatsoever on the target disorder (and, to keep the example simple, negligible side-effects). Meanwhile, P_4, the pharmacologically produced purified water. Imagine that the component *e* had some mild therapeutic benefit for sore throats, perhaps by increasing the body's hydration. For further simplicity, imagine further that patient expectations and dispensing practitioner attitudes have no effect on sore throats.

Other pill medicines for sore throats, whose characteristic features (analogous to *a*, *b*, and *c* in our example) do have beneficial effects, are also washed down with water. Imagine it is recognized and well-known that liquid with which the pill is swallowed has a mild therapeutic benefit. In the case of Superwater, a pharmacological agent responsible for the outcome, it might be mistaken to call this treatment a nonplacebo *simply because it is pharmacological and has some remedial effect.*

The stubborn objector could, of course, insist that the feature *e* has misclassified as incidental. Because the therapeutic benefit of the purified water can be pharmacologically explained *and* has nothing whatsoever to do with patient beliefs and expectations, it could be argued that it is a misuse of language to call the factor placebogenic. In response, recall that incidental features, to Grünbaum, are whatever treatment features are *not* characteristic. These will therefore include more than psychological factors that are typically associated with 'placebos'. Although Grünbaum's usage may be revisionist to some, it remains better than other conceptualizations for two reasons.

First it avoids the problems with the other conceptualizations stated above, including that of identifying any particular feature, such as sugar, as necessarily placebogenic. Second, Grünbaum's scheme makes more sense than any other when it comes to the design of placebo controls in clinical trials. The aim of a placebo control in a trial is to isolate the characteristic features and measure whether these features have any effect. To accomplish this, all non-characteristic features, whether they have to do with patient beliefs or not, must be controlled for. Surely the water with which pills are swallowed is the type of factor we would like to control for. The alternative would be to allow Superwater to be considered a nonplacebo, and even eligible for patenting and market approval, which seems strange, if not absurd.

In short, and contra Greenwood, pharmacological features, like sugar, may be difficult to classify as necessarily characteristic. Even the example Greenwood provides does not straightforwardly count against Grünbaum.

4.4.3. Mistaken identification of psychological factors with placebo factors.

Both Waring and Greenwood complain that Grünbaum fails to capture the intuition that placebos are exclusively to do with psychological rather than incidental factors. Indeed they both suggest replacing Grünbaum's definition of placebos with one that is more closely tied to psychological factors such as patient expectation and practitioner enthusiasm. Waring states "psychological factors such as a patient's expectations of benefit seem closer to what we intend by the placebo concept rather than remedial failure" (Waring 2003, p.14). Greenwood states that "we [might] have an instance of a placebo effect, according to Grünbaum's definition, *even though no part of the effect is produced by psychological factors such as therapist/doctor commitment or client/patient expectancy*" (Greenwood 1997, p. 499, emphasis original).

There are two ways to interpret these claims. The first is that placebos just are psychological factors and the second is that placebos must be identified with particular psychological factors, namely therapist/doctor enthusiasm and patient expectations.

The problem with identifying placebos with psychological factors is analogous to that of identifying placebos with sugar pills. Just as sugar pills are not placebos for all disorders, so psychological factors are not always placebos. If we accept Waring and Greenwood's suggestion, then we would rule out all psychological therapy as placebic *a priori*. Surely certain forms of behaviour therapy is (or at least could be) more than a placebo – at least, if it is not then this is an empirical issue and not one to be decided by

a priori definitional decree. Grünbaum is surely correct that what counts as a placebo must be relativized to a particular therapeutic theory and a particular disorder.

The objection also fails on the second interpretation for two reasons. For one, there are some cases where participant expectations and therapist commitments are not *necessarily* placebos. Imagine a type of psychological therapy called "You can do it!" Therapists trained in "You can do it!" have to be exceptionally committed and enthusiastic about their work or they lose their license. Their method involves getting the patients themselves to expect and believe that they will recover completely and quickly. Although empirical investigation may reveal that such a treatment is indeed what Waring or Greenwood would like to call placebos, surely we do not want to decide this *a priori.*

Greenwood and Waring could respond that being psychological was a necessary condition for something to count as a 'placebic factor' but not sufficient in itself. However, there is some evidence that the colour of a pill can have effects[50]. The property of being blue, red or yellow, is not psychological, nor does it have anything to do with patient/doctor interaction, yet it may be perfectly reasonable to consider colour to be placebogenic.

Further, although it is true that Grünbaum claims that *generic* placebos need not include any psychological factors. He is quite clear that both intentional and inadvertent placebos *usually* contain elements of practitioner and patient belief and enthusiasm (see definitions of inadvertent and intentional placebos above). Because the accusation that Grünbaum's placebos do not include patient expectation and therapist commitment has been raised in major journals twice, and I think it is mistaken, it is worth citing the other places where Grünbaum allows explicitly for these factors to be placebogenic:

> Note that each of the two species of placebo therapy I have considered is defined by a *conjunction* of two sorts of statements: (1) an assertion of *objective fact* as to the therapeutic failure of *t*'s characteristic constituents with respect to *D*; and (2) claims concerning the *beliefs* held by the therapist and/or the patient in regard to *t* (1986, p. 37).

Or, a few pages earlier,

> Research during the past three decades has envisioned ... that the therapeutic success of placebos may depend on certain kinds of characteristic of attitudes possessed by the treating physician. It should be

[50] See chapters 6, 7 for references and further examples.

> noted that my explications of both the intentional and inadvertent species of placebo have made provision for these two possibilities (1986, p. 35).

Grünbaum is unambiguous that patient beliefs and practitioner attitudes are factors that usually count as part of what it is to be a placebo. Like sugar, however, they are not necessarily placebogenic, as the "You can do it!" example illustrates.

In short, it seems that Waring and Greenwood's suggestion to identify placebos with psychological factors is mistaken. For one, neither general psychological factors nor participant expectations and practitioner enthusiasm need be placebo effects *simpliciter*. Further, they fail to recognize that placebos are either intentional or inadvertent, and in both cases patient belief and practitioner enthusiasm will usually play a role. To insist, however, that patient belief and practitioner enthusiasm are always necessarily placebogenic, ignores Grünbaum's main point, namely that *nothing is a placebo simpliciter*. Claiming that some particular factor, even patient/practitioner belief is *necessarily* placebogenic will fall prey to the same reductio as regarding the sugar pill as a necessary placebo, as the "You can do it!" example suggests.

4.4.4. Grünbaum's apparent failure to place suitable restrictions on what counts as the relevant therapeutic theory

Hróbjartsson acknowledges that Grünbaum 's is the best definition of placebos offer, but claims that even Grünbaum 's conceptualization is unsatisfactory because it offers no guidance as to what counts as a justifiable therapeutic theory:

> Grünbaum's definition is by far the best proposal for a formal definition of placebo. Despite that, the definition is not generally regarded a satisfying conceptual clarification, partly because he does not give any clear criteria for what constitutes a good therapeutic theory, nor what to do when two [such] theories compete (Hróbjartsson and Gøtzsche 2001).

With no guidance about how to construct good therapeutic theories, we may not be much closer to a useful guide to conceptualizing placebos than we were before.

In this section I will argue that Hróbjartsson's objection points to a real underlying problem with Grünbaum's scheme but that the problem is inevitable on any scheme and that Grünbaum's definitions retain their advantages over other proposals.

For example, consider the example I used above to illustrate Greenwood's earlier objection, where the only remedial feature, an antibiotic, was classified by the current therapeutic theory as incidental for the treatment of bacterial pneumonia. My

response was that *if* we knew that antibiotics were remedial for bacterial pneumonia (drugs don't just come with 'antibiotic' written on them), then the very theory which classified the antibiotic as incidental *should have* classified it as characteristic.

To be sure, the example of antibiotics is loaded - 'antibiotic' is a heavily therapeutic theory-laden term. A better example might be to replace 'antibiotic' with 'substance X'. In this example, 'substance X' would be correctly identified as characteristic, but some accepted theory failed to identify it as such. In a real example, olive oil was once used in placebo capsules for trials of cholesterol-lowering agents before there was evidence that olive oil reduced cholesterol (Golomb 1995). Although olive oil was not considered characteristic by the therapeutic theory at the time, it may have had effects nonetheless. The therapeutic theory, in the case of substance X and (in the past), olive oil, failed to correctly identify the characteristic features.

In response, of course therapeutic theories are not omniscient, and can be mistaken. Grünbaum acknowledges that the therapeutic theory ψ can be ignorant of the effects of certain characteristic and incidental features: "if some of the incidental constituents of t are remedial but presently elude the grasp of ψ, the current inability of ψ to pick them out from the treatment process hardly lessens the objective specificity of their identity, mode of action, or efficacy" (1986, p. 33). Grünbaum need merely add that in practice, some of the unknown incidental factors would be better described, by a truer theory, as characteristic. This is not a problem with Grünbaum's scheme *per se* but a problem with the fallibility of science in general –theories will not be completely correct all the time.

Further, a word must be said about the scope of the problem with mistaken therapeutic theories. It is unlikely that there are a great number of cases such as the imaginary example of substance X, or the one Golomb describes. In fact, in the case of pharmaceutical drugs and most herbal treatments, speaking of a therapeutic theory is arguably redundant. The characteristic feature is the patented chemical or the herb, and the incidental features are whatever else is involved in the treatment.

However, in other cases, the characteristic/non-characteristic divide is not so clear, and the possibility of different therapeutic theories leading to different 'placebos' is possible, an example being both supervised flexibility (Dunn et al. 2002) and supervised relaxation (McCann and Holmes 1984) used as 'placebo' controls for exercise. Similarly, several different treatments have been used as 'placebos' for

acupuncture therapy. One involves a needle that does not penetrate the skin (Streitberger and Kleinhenz 1998; Kaptchuk et al. 2006), while the other penetrates the skin but not at the alleged 'acupuncture points' (Haake et al. 2007) (I will describe these in greater detail in the next chapter).

In these cases the underlying therapeutic theory may be insufficiently 'strong' to distinguish between the characteristic and non-characteristic features. In these cases there is room for different (usually, as we shall see next chapter, implied) therapeutic theories to lead the (again, usually implied) design of different placebos.

However in the cases where the underlying therapeutic theory is not sufficiently well established, the problem is not that there are conflicting therapeutic theories, but is the more basic problem that current science has not identified the causal pathways of all the treatment features and the characteristic features cannot be identified. In these cases, until the therapeutic theory becomes better established, it is surely better to avoid the design of placebos at all and test their efficacy in 'active' controlled trials. Yet even here, Grünbaum's scheme retains its advantage because it identifies the problem with placebos for these treatments.

More importantly, Hróbjartsson's alternative to Grünbaum is even more problematic. Hróbjartsson suggests that, we change our focus from a conceptual one to a 'practical' one, and define placebos "practically as an intervention labelled as such in the report of a clinical trial" (Hróbjartsson and Gøtzsche 2001, p. 1595). Although clinical researchers are, by and large, unlikely to err in their design of placebos, this 'practical' solution has, at best, a very thin normative foundation. In fact, Hróbjartsson jumbles (along with placebo pills and injections) relaxation (described as a placebo in some studies and a treatment in others), leisure reading, answering questions about hobbies, newspapers, magazines, favourite foods and sports teams, talking about daily events, family activities, football, vacation activities, pets, hobbies, books, movies, and television shows as placebos (Kirsch 2002).

Further, one of Hróbjartsson's problems with attempts to conceptualize 'the' placebo is that it fails to take into account the diversity of treatments that can be called placebos. "It might be time to stop using the term placebo effect and instead specify which kind of intervention one is referring to" (Hróbjartsson 2002). Indeed given that different placebos have different effects, Hróbjartsson's suggestion to avoid talk of placebos is very useful. Imagine that trials of a pharmaceutical analgesic drug consistently demonstrates 25% effects over and above 'the' placebo effect. Then,

imagine that acupuncture as an analgesic consistently demonstrates 10% superiority to 'the' placebo effect. With this limited information we would be correct to conclude that the pharmaceutical analgesic was more effective than acupuncture. However what if we subsequently learned that 'acupuncture placebo' was 50% more effective than 'pharmaceutical drug placebo'? With this updated information, we might infer that acupuncture was *more* effective than the pharmaceutical drug[51]. If we abandoned talk of 'the' placebo, as Hróbjartsson suggests, then the mistaken inferences about comparative effectiveness would vanish. Indeed whether a placebo is red, blue, yellow, or green, whether it is injected or swallowed, and whether it has a brand name can all affect its effectiveness[52]. Hróbjartsson's failure to consider Grünbaum's scheme in details makes him oblivious to consider how the scheme solves many of the problems he himself posits.

Ironically, Grünbaum's conceptualization itself solves the apparent problem that conceptualizations obscure the potential differences between the treatments we call 'placebos'. Relativizing the incidental and characteristic features (and hence what counts as a placebo) to a therapeutic theory means that treatments characterized as placebos for a target disorder by one therapeutic theory will not necessarily be the same (indeed they will usually be different) treatments characterized as placebos by a different therapeutic theory. Hence, one of the issues Hróbjartsson has with attempts to conceptualize the placebo seems to be solved by Grünbaum's scheme.

In sum, Hróbjartsson is correct that an adequate definition of the placebo has proven especially hard to come by. However, he fails to consider Grünbaum's scheme in any detail. Closer examination reveals not only that Grünbaum's conceptualization solves some of the Hróbjartsson's problems, but also that the problem with failure to place strictures on therapeutic theories might be limited in scope.

4.5. Conclusion and Implications for Placebo Controls

Grünbaum's definition, in the modified form I proposed, resists the common objections and overcomes the generally confusing attempts to define the placebo.

---

[51] This example is taken from a real study which I will discuss in more detail next chapter (Haake et al. 2007).

[52] See chapters 6, 7 for references and further examples.

The modified version of Grünbaum's avoids the ambiguity inherent in the common term 'specific', and at the same time provides us which a scheme for classifying treatments as placebos and nonplacebos in a wide variety of cases: a treatment *t* is a nonplacebo if its *characteristic features* have effects, and it is a placebo if its characteristic features do not have effects on the target disorder (and if the overall effects are not toxic). Relativizing the concept of placebo to a particular therapeutic theory prevents us from calling anything a placebo *simpliciter*, as the example of the sugar pill for diabetics demonstrates.

Waring's objection that Grünbaum's scheme allows us to characterize toxic agents as placebos can be absorbed by the modified version of Grünbaum's scheme where, in addition to *placebos* and *nonplacebos*, we introduce *toxic agents* and *nocebos*. Greenwood's objection that pharmacological factors are necessarily nonplacebic fails for the same reason that calling sugar (which could also be pharmacological) a placebo factor does – sugar is not a placebo for the treatment of diabetes. Similar problems plague the insistence that psychological factors must be placebic.

Hróbjartsson's complaint that Grünbaum offers no guidance as to what counts as a legitimate therapeutic theory is grounded in a failure to think through the consequences of Grünbaum's scheme carefully. Further, Hróbjartsson's suggestion that we avoid talk of 'the' placebo effect is implied by the very point about Grünbaum's scheme that Hróbjartsson takes issue with, namely the relativization of placebo to a therapeutic theory.

In order to use 'scientific common sense' to evaluate whether placebo controlled trials are superior to 'active' controlled trials and (although this is less obvious but will become clear over the course of the next two chapters), to evaluate whether double blinding is always of methodological value, we need to know what placebos and placebo controls are. With an adequate conceptualization of placebos in hand, I will proceed to investigate what counts as a 'legitimate' placebo in the next chapter.

5. Chapter Five. Placebo Controls: Problematic and Misleading 'Baseline Measure of Effectiveness'

> ***Doctor (to patient):*** *First, take the white pill with a glass of water, then take the yellow pill with a glass of water, then take the red pill with a glass of water.*
>
> ***Patient:*** *What on earth do I need so many pills for?*
>
> ***Doctor:*** *You don't drink enough water.*
>
> *- Anonymous joke*

> *Is there a need to control the placebo in placebo controlled trials?*
>
> *- (de Craen, Tijssen, and Kleijnen 1997)*

5.1. Setting Standards for Legitimate Placebo Controls

It is often alleged that placebo controlled trials are methodologically superior to 'active' controlled trials:

> Unfortunately, ACETs ('active' controlled trials] are often uninformative. They can neither demonstrate the effectiveness of a new agent nor provide a valid comparison to control therapy unless assay sensitivity can be assured, which often cannot be accomplished without inclusion of a concurrent placebo control group (Temple and Ellenberg 2000, p.462).

Similar views can be found (ICH 2000; Gomberg-Maitland, Frison, and Halperin 2003; Hwang and Morikawa 1999). However in order to evaluate this argument, we need to decide what counts as a placebo control. Because there has been no acceptable definition of placebos *per se*, this task has not yet been completed.

Placebo control treatments are the standard by which the effectiveness of experimental treatments are often gauged. A placebo controlled trial provides a rough estimate of the magnitude of the effects of the characteristic features of the experimental treatment (i.e. fluoxetine in treatment with a Prozac pill). This estimate is obtained by subtracting the average outcome in the placebo control group (who are treated with, say, sugar pills) from the average outcome in the experimental group (who are treated with, say, real Prozac pills). Failure to demonstrate superiority to placebo control is taken to indicate that the characteristic features of the treatment are ineffective, and vice versa.

In spite of the vital role of placebo controls in determining whether the characteristic features of experimental treatments are effective, standards for what counts as *legitimate* placebo controls have not been made explicit. Hróbjartsson and Gøtzsche, two prominent investigators of placebo effects circumvented the hard work of defining placebos and placebo controls (claiming that current definitions are "illogical") and define placebos "practically as an intervention labelled as such in the report of a clinical trial" (Hróbjartsson and Gøtzsche 2001, p. 1595)[53]. But allowing any treatment used as a placebo control in a clinical trial invites inaccurate estimations of the magnitude of the effects of the characteristic features.

Even Grünbaum, whose (1986) conceptualization of *placebos* was, as I argued last chapter, defensible with some modification, his definition of placebo *controls* is highly problematic. Grünbaum's conceptualization of placebos can be summarized very briefly as follows: if a treatment's characteristic features (i.e. fluoxetine in treatment for depression) are effective for the target disorder, then it is a nonplacebo; otherwise it is a placebo (Grünbaum 1986`, p. 22-28). Grünbaum then claimed that the purpose of placebo *controls* is to separate the effects of the characteristic features from the potential effects of expectancy:

> Turning now to placebo *controls*, we must bear in mind that to assess the remedial merits of a given therapy ... it is imperative to disentangle from each other two sorts of possible positive effects as follows: (1) those desired effects on *D* [the target disorder, i.e. depression], if any, actually wrought by the characteristic factors ... ; and (2) improvements produced by the expectations aroused in both the doctor and the patient by their belief in the therapeutic efficacy of ... [the therapy]. To achieve just such a disentanglement, the baseline measure (2) of expectancy effect can be furnished by using a generic placebo *t* in a control group of persons suffering from *D* (Grünbaum 1986`, p. 26)

In light of these considerations, Grünbaum defines a placebo control as follows:

> A treatment type *t* functions as a 'placebo control' ... just when the following requirements are jointly satisfied: (1) "*t* is a *generic placebo* [i.e. it has no remedial characteristic features] for *D* and (2) "the experimental

---

[53] However, when they discuss their methods, they admit to omitting trials where "it was very likely that the alleged placebo had a clinical benefit not associated with the ritual alone (e.g. movement techniques for postoperative pain)" (Hróbjartsson et al. 2007, p. 1595). This suggests that, although they refrain from defining the placebo explicitly, they do have *some* definition in mind. I discuss this issue in more detail in the next chapter.

> investigator ... believes that *t* is not only a generic placebo for *D*, but also is generally quite harmless" (Grünbaum 1986`, p. 26).

However, quite trivially, the 'baseline measure' of expectancy effect is *not* furnished by using a placebo *á la* Grünbaum; other factors such as spontaneous remission (many disorders go away without being treated at all) could explain the response in the placebo control group, even if expectancy had no effects at all. Waring (2003) makes this point succinctly:

> It [Grünbaum's definition] also ignores the factor that might explain the response in a placebo control group, such as the natural course of the illness (e.g., spontaneous remission, regression to the mean). These factors relate to the disorder and are not incidental to a treatment (Waring 2003`, p.9).

A placebo does not furnish a baseline measure of expectancy, but rather a baseline measure of all sources of non-characteristic effects including, but not limited to, expectancy.

In this chapter I will defend the following definition:

Legitimate placebo control: a treatment *t'* is a *legitimate placebo control* in a trial of test treatment *t* for a target disorder *D* if and only it satisfies the following two conditions:

(1) The treatment process *t'* (the 'placebo' control) must contain all the non-characteristic features (defined by the therapeutic theory ψ) of the test treatment *t*;

(2) The treatment process *t'* (the 'placebo' control) has no additional features over and above the non-characteristic features (defined by the therapeutic theory ψ) of the experimental treatment.

My proposed definition of legitimate placebo controls borrows several terms from Grünbaum's definition[54] of placebos that I will explain briefly with the help of an example.

---

[54] Grünbaum argued, successfully I think, that previous descriptions of placebos as treatments with 'nonspecific' effects (Shapiro and Morris 1978), prescribed for the psychological benefit to the patient rather than for any 'specific' effect, is fundamentally flawed. Placebos can have very specific effects.

Take the treatment process involving fluoxetine (the formerly patented chemical in a Prozac pill) for depression. This treatment process could involve:

- Fluoxetine
- The other ingredients of the Prozac pill (pill casing, etc.)
- (Sometimes) the water with which the pill is swallowed
- (At least where Prozac is only available by prescription) a consultation with a (hopefully) sympathetic physician or psychiatrist
- The participant's expectation that he or she is being treated with fluoxetine

In the treatment process described above, the only 'characteristic' feature is fluoxetine; the other features are 'non-characteristic'[55]. The designation of characteristic features is relative to both a target disorder and a therapeutic theory. A sugar pill, for instance, may be a placebo for the treatment of headaches but is hardly a placebo for diabetes. Similarly, certain forms of therapist or doctor commitment could be incidental for surgery but characteristic for certain types of psychological therapy. This is not to claim that these characteristic features for a psychological treatment are in fact effective, but merely to avoid *a priori* classifications of all psychological treatments (or treatment with sugar pills) as placebos. For the purposes of the present discussion, the therapeutic theory will be nothing more than a specification of characteristic features. In the case of therapies involving drugs, such a classification is straightforward – it is simply the chemicals such as fluoxetine. However, as we will see below, the classification is not so straightforward for non-drug treatments such as those involving exercise. The other

---

> Mindful of the potential side effects of tranquilizers and analgesics, the doctor decides to employ a little benign deceit and gives B a few lactose pills, without disabusing B of his or her evident belief that he or she is receiving a physician's sample of analgesics. Posit that shortly after B takes the first of these sugar pills, the headache disappears altogether. Assume further that B's headache would not have disappeared just then from mere internal causes. ... Thus B assumedly received the same headache relief from the mere sugar pill as he or she would have received if a pharmacologically noninert drug had been slipped into his food without his knowledge.
>
> Clearly, in some situations, the therapeutic effect of the sugar pill placebo on the headache can have attributes fully as sharply defined or 'specific' as the effect that would have been produced by a so-called 'active' drug .... Moreover, this placebogenic effect can be just as precisely described or known as the nonplaceboegenic effect of aspirin. In either case, the effect is complete headache relief" (Grünbaum 1986`, p. 31).

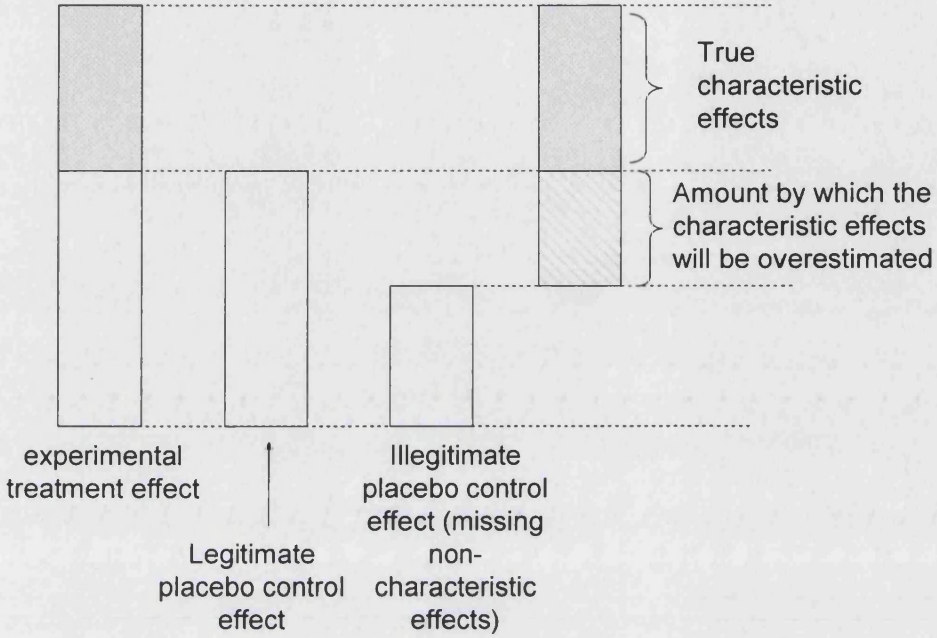
[55] Grünbaum would describe the other features as 'incidental'. I choose 'non-characteristic' for simplicity and accuracy – the categories, even to Grünbaum, are exclusive and exhaustive.

features, which will sometimes be difficult to specify completely, will be non-characteristic.

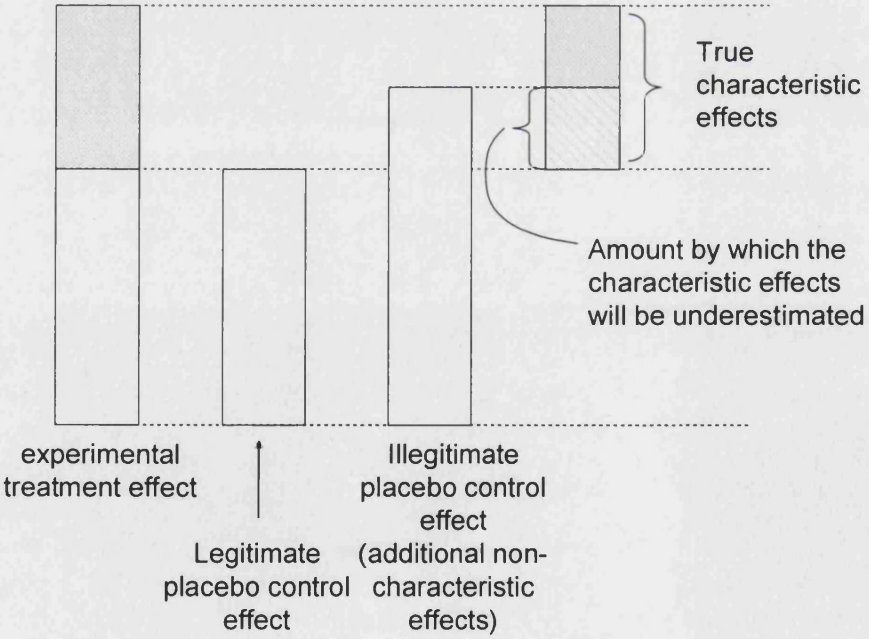
To anticipate, I will argue that in practice, many, perhaps even most, actual treatments with placebo controls fail to satisfy at least one of the two conditions for legitimacy. The danger with illegitimate placebo control treatments is difficult to overstate. If treatment with the 'placebo' control is less effective than the non-characteristic features of the experimental treatment, then the characteristic effects will be erroneously exaggerated, and vice versa. The pitfalls of illegitimate placebo control treatments situation is best understood with the aid of a diagram (see chart below).

4. Chart 5.1: Illegitimate Placebo Controls Deliver Mistaken Estimates of Effect Size

(a) PCT employing an illegitimate (insufficiently effective) placebo control



(b) PCT employing an illegitimate (overly effective) placebo control



As is clear from the chart, if treatment with the 'placebo' control is more effective than treatment with a legitimate placebo, the effectiveness of the characteristic features will be over-estimated (see chart 5.1 (a)). On the other hand, if treatment with a 'placebo' control is less effective than treatment with a legitimate placebo, the effectiveness of the characteristic features will be under-estimated (see chart 5.1 (b)).

I will now describe the most common ways placebo controls fail to be legitimate.

5.2. Failure to Satisfy the First Condition: Placebo Controls that are 'Missing' Non-Characteristic Features of the Experimental Treatment.

The first condition a control treatment must meet in order to be considered legitimate is that it must contain all the non-characteristic features of the experimental treatment. To recap:

(1) The treatment process *t'* (the 'placebo' control) must contain all the non-characteristic features (defined by the therapeutic theory ψ) of the test treatment *t*

Failure to perform the first function will lead to an overestimation of the characteristic effects, and is illustrated in column B of the chart above.

The first condition appears to be easily met in drug trials. Here the placebo control treatment must merely involve everything *but* the chemical (i.e. fluoxetine in a Prozac pill). This is achieved by treating participants with pills that look like the 'real' pills but that do not contain the experimental chemical. The chemical is replaced with an innocent substance such as lactose. Such a placebo control has an additional feature, i.e. lactose, but it does not appear to be missing anything.

Yet, as I will illustrate by describing a flaw in Grünbaum's definition, in practice many placebo control treatments might be missing the feature of participant belief or expectancy that they are being treated with a non-placebo. A generic placebo, in Grünbaum's view, is defined 'negatively' as a treatment that has no characteristic features for a particular disorder. There is no stipulation that it must contain *all* the incidental features of the test treatment – indeed Grünbaum acknowledged that these will be difficult, if not impossible to specify. However if the control treatment process does not include participant knowledge that they are (or could) be receiving the experimental intervention, then any difference between the effectiveness of the control

and test treatments may be due to the difference in *expectancy* and not the difference in characteristic features.

For example, reconsider the imaginary placebo controlled trial of Prozac therapy for depression. Imagine that the experimental group received pills that contained fluoxetine (the only characteristic feature) with "PROZAC" written on them, while the control group received sugar pills with the words "PLACEBO" written on them. Treatment with the "PLACEBO" pills would count as placebos on Grünbaum's scheme since they did not contain fluoxetine. Yet, since the participants in the trial would identify the placebo control treatments as placebos, they would have lower expectations regarding recovery compared to the participants receiving Prozac. The different levels of expectation, and not any potential effects of fluoxetine could explain the difference between treatment with 'real' versus 'placebo' Prozac therapy. In short, the control treatment is missing the expectation of being treated with fluoxetine.

The easy way to make sure that expectations are the same in the experimental and control group (and hence that there are no 'missing' expectations in the 'placebo' group) is to blind (or 'mask') the participants. Although masking might sound easy to achieve (and is often talked about in the medical literature as if it were), a closer look reveals that in practice it is extremely difficult. This is partly due to the fact that trials must be performed with full informed consent of participants, whereby participants are supposed to be informed of the fact that they are enrolled in a placebo controlled trial, and of the known potential side effects of the experimental treatments. This means that to successfully mask an informed participant, the control must be sufficiently similar to the experimental treatment process in several ways.

The first and most interesting similarity that the placebo control must have with respect to the experimental treatment is effectiveness: treatments that contain dramatically effective characteristic features cannot be tested in double masked trials. If I were enrolled in a placebo controlled trial of a new drug to treat severe colds, and my severe cold symptoms completely disappeared within seconds of taking the drug, then I would deduce that I had been given the test treatment and not the placebo. The problem is that it may be difficult to distinguish between trials which come unblind due to the effectiveness of the test treatment or because of other reasons, such as expectations of investigators administering the intervention. The fact that certain treatments may have such dramatic effects that the blind cannot be maintained highlights the problem with regarding double-blind trials as best evidence: treatments with the most dramatic effects

can't be supported by 'best evidence'! This alone is surely grounds to question the claim that masking the participants and dispensers is a universal virtue of medical studies[56].

Second, the control and test treatments must appear similar to the senses: they must look, smell, and taste similar to the test treatment. These superficial features of the treatment process are non-characteristic, and although not necessarily effective in and of themselves, may be sufficient to unmask the study if missing from the 'placebo' control. If, for example, the control pill tastes like candy while the experimental treatment, say vitamin C, pill tastes sour, strong suspicions amongst the participants about which group they are in might arise. Similar arguments can be made about the necessity of having the test and control treatments feel and smell the same.

Third, and more significantly since it has tended to be underplayed until very recently, the *side-effects* of the experimental and control treatments must be similar. If, for example, dry mouth is a common side-effect of an antidepressant drug, and an informed participant in an antidepressant placebo controlled trial gets a dry mouth, then that participant may well suspect that she was taking the experimental drug and not the placebo. Likewise, a participant in the same trial who did not experience any side effects might suspect that she was in the placebo group. This would mean that although the trial was blinded at the very outset, the blind was soon broken.

As has been recognised more recently (Kirsch and Sapirstein 1998; Kirsch 2002; Moncrieff and Wessely 1998; Moncrieff and Kirsch 2005; Moncrieff 2003; Edward et al. 2005) in order to eliminate this possibility, 'active placebos', which are not only sensibly indistinguishable from the test treatment, but also imitate its side-effects need to be employed. In the case of the antidepressant with the side effect of a dry mouth, the placebo control might include some agent that is believed to have no characteristic effect on the target disorder of depression but does cause dry mouth. Making placebos

---

[56] To be sure, masking the outcome assessors (who are often the same people as the dispensers), statisticians, and even manuscript writers, is not only generally easier to achieve than masking of the participants and dispensers, but may even rule out real confounders. See chapter 5 for detailed discussion of the value of double masking.

active is no easy feat, and it is questionable whether it is ever fully (or even near-fully) possible in practice[57].

It does not follow, of course, from the fact that active placebos are difficult to construct, that we should abandon the use of placebo controls – perhaps we should simply try harder to make our placebos active. It does suggest, however, that in cases where it is unlikely that our attempts to design active placebo controls will *fail*, that we should consider other methods, such as 'active' controlled trials. (See chapters seven and eight for a critical evaluation of the arguments that 'placebo' controlled trials are methodologically superior to 'active' controlled trials.)

In addition to the similarities that must hold between the experimental and control treatments, blinding of the dispensing physician is a necessary condition for the retention of a strict single blind. A well informed clinician will be especially good at detecting features, effects, and side effects of the test treatment. The administering investigator's knowledge of which treatment each participant is being given can translate to participant knowledge. The investigators could either inform the participants explicitly or implicitly via 'subtle cues' which group they are in.

Needless to say, maintaining successful participant masking is difficult to achieve. The 'placebo' control for an intervention whose side-effects that are not easily imitated, whose effects are dramatic, or whose characteristic features have distinctive smells or tastes will not induce the same participant or dispensing physician expectations. Treatment with a 'placebo' control that does not succeed in inducing the same expectations regarding recovery as treatment with the experimental treatment will be missing any effects of these expectations. If these expectations have effects, and there is no doubt that in at least some cases they do, then the 'placebo' control treatment that does not include these expectations will not be legitimate.

Although worrisome even for drug trials, failure to satisfy the first condition is amplified for non-drug treatments. For example, although placebo controlled trials of surgical techniques have been performed(Freed et al. 2001; Moseley et al. 2002; Connolly et al. 2003; Olanow et al. 2003; Gragoudas et al. 2004), they are generally considered unethical, mostly because of the inherent risks involved even with 'placebo' surgery (Heckerling 2006). Yet, even if considered ethical, the surgeon performing the

---

[57] I discuss the virtues and vices of keeping a trial double blind more carefully in the next chapter.

sham operation will (one hopes) be aware of whether he is performing real or sham surgery, which makes the attitudes of the surgeon potential confounders. This undermines the justifiability of the placebo control used in surgical trials. The only conceivable way to keep surgical trials double masked would be if the procedure to be entirely mechanized. Although robots sometimes assist physicians (Katz et al. 2006; Worn 2006; Diks et al. 2007; Suzuki et al. 2007), they aren't yet sophisticated enough to perform operations on their own. In addition to surgery, psychotherapy (or any form of 'talking' therapy), and acupuncture cannot (at least not straightforwardly) be imitated by treatments that permit the trial to remain double-blind.

Of course, if expectations (or any missing feature for that matter) have no effects on the target disorder, then its absence will not undermine a control treatment's legitimacy. In some cases, expectations might not have effects. Also, even if effective, if expectations have very small effects relative to the effects of the characteristic features, then any difference in expectancy may not be important. For example, patient expectation of receiving an appendectomy for acute appendicitis could have effects, but the effects are likely to be unimportant relative to the characteristic effectiveness of appendectomies. Non-characteristic features with no effects, then, can be missing from legitimate placebo controls.

It must be remembered, however, that even if a feature has no effects on the target disorder, it could still undermine legitimacy by causing the unmasking of the participants. For example, if the 'placebo' control of an antidepressant trial was missing the side-effect of dry mouth, and participants correctly guessed that they were in the 'placebo' group based on the failure of the treatment to produce this side effect, then the missing feature, although ineffective for the target disorder, threatens the legitimacy of the 'placebo' control treatment. Here again, however, in order for the legitimacy to be threatened by the missing feature that causes unmasking, there must be important expectancy effects.

With this in mind, the first condition for legitimacy can be revised to specify that only *relevant* non-characteristic features cannot be missing from the control treatment. A relevant non-characteristic feature is one that (a) has or might have important effects, and (b) (where the effects of expectations have important effects) does not cause the unmasking of the participants. The first condition for legitimacy is then:

(1') The treatment process *t'* (the 'placebo' control) must contain all *relevant* non-characteristic features (defined by the therapeutic theory ψ) of the test treatment *t*

I will now proceed to examine the second condition for legitimacy in more detail.

5.3. Failure of a Placebo Control to Satisfy the Condition of Not Containing Additional Non-Characteristic Features

Recall the second function a treatment process must perform in order to be a legitimate placebo control treatment:

(2) The treatment process *t'* (the 'placebo' control) has no additional features over and above the non-characteristic features (defined by the therapeutic theory ψ) of the experimental treatment.

Failure to satisfy this condition, i.e. if the 'placebo' control therapy has additional features over and above the non-characteristic features of the experimental therapy, the characteristic effectiveness will tend to be underestimated. This is illustrated in column C of the chart above.

Grünbaum's definition of placebo control fails to satisfy this condition. At least in principle, Grünbaum's definition does not rule out a treatment that is a nonplacebo according to a *different* therapeutic theory as a placebo control. If a therapeutic theory considered fluoxetine to be the only characteristic feature of treatment with Prozac, treatment involving amitriptyline (the characteristic feature of treatment with older tricyclic antidepressants such as Tryptanol) could be a placebo control treatment according to a this reading of Grünbaum.

Picking up on this weakness, Greenwood states:

> The problem with this purely negative characterization of a placebo control treatment is that *all* alternative forms of professional psychotherapy, with alternative theoretical justifications, will count as placebo control treatments with respect to any particular experimentally evaluated form of professional psychotherapy. Yet such alternative treatments cannot function as effective placebo control treatments, for they do not enable us to assess whether the efficacy of any professional-experimental treatment is due to the presence of theoretically specified factors or the presence of naturalistic factors such as client expectancy and therapist commitment common to experimental and placebo treatments. For in this case, if the recovery rates for the professional-experimental treatment are not significantly superior to those of the placebo control treatment, this may be because of the naturalistic factors common to both treatments or because of the *additional* theoretically

> justified factors present in the alternative professional treatment. For this reason, if no other, we ought not to completely abandon the original psychological reference in our definition of a placebo treatment, which, it is suggested, may be best defined as any treatment that lacks the theoretically efficacious elements of a particular experimentally evaluated treatment (in medicine or psychotherapy), and has no known theoretical justification *over and above* its justification in terms of naturalistic factors such as client expectancy and therapist commitment (Greenwood 1996`, p. 615).

Greenwood objects that according to Grünbaum's scheme, one type of psychotherapy that has characteristic effects according to one therapeutic theory may well count as a placebo for a different type of psychotherapy. Or, amitriptyline could be classified as placebogenic as far as treatment with Prozac is concerned.

However, at least in the latter case it would be a strange therapeutic theory about treatment with Prozac indeed that did not include statements about other known antidepressant agents. Clearly no one could think of Grünbaum's criteria for placebos (that they do not contain characteristic features as specified by a particular therapeutic theory) as anything other than a necessary rather than a sufficient condition – no one is happily going to regard amitriptyline, for example, as non-characteristic for depression. Still, because of the importance of spelling this out clearly, Greenwood's criticism, at least as a call for clarification, is fair.

Nonetheless, cases similar to the imaginary example of amitriptyline being used in a placebo treatment for a Prozac trial may sometimes occur when we are genuinely ignorant of the properties of the agent in question. The characteristic chemical of the experimental treatment will usually have to be replaced with some other substance, a 'replacement substance'. For instance, fluoxetine will be replaced by a substance such as lactose for the Prozac placebo. The replacement substance is an additional feature, but because it is supposed to have no effects on the target disorder, it will not render the placebo control therapy illegitimate. It is conceivable that apparently innocent replacement substances (such as lactose) in fact have effects on the target disorder being studied. Indeed, in a real example, olive oil was used as a replacement substance for trials of cholesterol-lowering agents before there was evidence that olive oil reduced cholesterol:

> several early papers exploring the use of cholesterol-lowering agents to curb heart disease did in fact name the placebos used: olive oil in one case, and corn oil in another. Mono- and poly-unsaturates such as olive oil and corn oil are now widely known to decrease low-density lipoproteins, so that with hindsight these agents may not have been inert with respect to the outcome

studied. Indeed, it was noted in one such study that the rate of cardiac mortality was lower in the placebo group than expected (Golomb 1995).

Although olive oil was not considered characteristic by the therapeutic theory at the time, it was subsequently discovered to have cholesterol lowering properties. To be sure, the amount of olive oil used in the pills may not have been sufficient to produce an effect, but the example illustrates how even simple pill placebos may fail to perform their required function.

In the worst case the 'double positive paradox' will result. Although the experimental intervention may be no more effective than the 'placebo' control, both the 'placebo' control and the experimental treatment are, from a more fundamental point of view so effective that it seems to be a misuse of language to call them 'placebos'. To take a modified real example that I will consider in more detail below, treatment with real acupuncture is often no more effective than treatment with 'placebo' acupuncture, but both real and 'placebo' acupuncture are, at least in some trials, more effective than conventional therapy (which included the administration of analgesic drugs).

Thinking about replacement substances raises the necessity of modifying the second condition in a similar way to the first. Some additional non-characteristic features may not have effects on the target disorder, or they may be unimportantly effective relative to the characteristic features (i.e. they may not cause unmasking). For example, lactose is innocent for most (but not all) target disorders. Again, even if the additional feature does not have *direct* effects on the target disorder, they could, in some cases, lead to the unmasking of the participants. For example, if the 'placebo' control of an antidepressant trial had the additional feature of sugar, and this led to participants' correct guesses that they were in the 'placebo' group because the additional feature made the pill taste sweet, then the additional feature threatens the legitimacy of the 'placebo' control treatment. Here again, however, in order for the legitimacy to be threatened, there must be important expectancy effects.

With this in mind, the second condition for legitimacy can be revised to specify that only *relevant* non-characteristic features cannot be added from the control treatment. Like before, relevant non-characteristic feature is one that (a) has or might have important effects, and (b) (where the effects of expectations have important effects) does not cause the unmasking of the participants. With that in mind, a more precise second condition can be stated as follows:

(2') The treatment process *t'* has no additional features over and above the non-characteristic features of the experimental treatment.

To sum up what has been said thus far, treatment with legitimate placebo controls must be exactly as effective as treatment with the non-characteristic features of the experimental treatment. The easiest way to achieve this is, of course, for the control treatment to include all and only the non-characteristic features of the experimental treatment. In the relatively simple case of treatment with drug placebos, where the characteristic chemical is 'removed' and replaced with a replacement substance such as lactose, placebo control treatments may appear easy to design. Yet it may be difficult to maintain the masking of participants in placebo controlled drug trials. If the masking is not maintained, then one source of non-characteristic effects of the experimental treatment, namely the effects of expecting to be treated with the experimental intervention, will be absent from the placebo therapy effects, and the first function required for legitimacy will not be performed. Then (although this is surely rare), apparently innocent 'replacement substances' could have characteristic effects for the target disorder. In short, close inspection reveals that even for drugs where the design of placebo controls appears straightforward, legitimate placebo control treatments may not be as common as one would like to believe.

These problems are only amplified with non-drug treatments, which I will now consider in more detail.

5.4. The Difficulty of Designing Legitimate Placebo Controls where the Characteristic / Incidental Divide is Difficult to Draw: An Illustration with Exercise

The design of legitimate placebo control treatments is only amplified for more complex interventions. The characteristic features of most drugs are not only easy to identify, but also easy to separate from the non-characteristic features. Metaphorically, we simply remove the 'inside' of the pill (the characteristic chemical compound), replace it with an ineffective substance, and leave the 'outside' the same. This simple technique is not available to many treatments, either because the characteristic features cannot be separated from the other features, or because the choice of which features are characteristic is controversial. I will illustrate the former problem with the example of exercise-therapy for depression.

is intended to imitate. They do not contain all the non-characteristic features of exercise, but are actually distinct treatments. Indeed supervised relaxation or flexibility are so different from exercise proper that to call them 'placebo exercise' or even 'sham exercise' seems incorrect. I will call control treatments whose very nature is different from the experimental treatment 'surrogate placebo controls'.

On the face of it one might be tempted to rule that surrogate placebos are necessarily illegitimate. They contain many features that the experimental does not have, and at the same time they are 'missing' many non-characteristic features of exercise-therapy. Treatment involving supervised flexibility, for example, is missing increased heart rate, which is an incidental feature of exercise-therapy as I have construed it. At the same time treatment involving supervised flexibility has the feature of extending muscle length, which exercise therapy does not have. Because they are both 'missing' and have additional non-characteristic features, we might suspect that surrogate placebo controls fail to perform both the required functions of a legitimate placebo control.

One could argue that a placebo control treatment can be legitimate even if it fails to satisfy my conditions for legitimacy so long as its *effects* are the same as the effects of the non-characteristic features of the experimental treatment. After all, the implicit rationale for the conditions for legitimacy is that the *effects* of the control treatment process must be the same as the effects of the non-characteristic features of the control treatment process. However it is difficult to specify what the *effects* of the non-characteristic features are in advance without measuring them directly, i.e. by employing a legitimate placebo control.

In addition, there are some specific non-characteristic features of exercise that supervised flexibility is missing, namely expectation that one is being treated with exercise. I will illustrate this with an example. In a recent study, low and high 'doses' of exercise (each spread out over 3 or 5 sessions) were compared with exercise 'placebo' for the treatment of major depressive disorder (Dunn et al. 2005). The investigators randomized 80 participants to receive one of 5 treatments for 12 weeks (see table below).

5. Table 5.1: Description of Exercise and 'placebo' Exercise Treatments in Dunn *et al.* 2005.

Name	Description	*Total* kilocalories per kilogram per week burned (kcal / kg / week)

Assume that the features (both incidental and characteristic) of an extended programme of exercise for the treatment of depression includes the following factors:

(1) belief that one is being treated with exercise
(2) participant / investigator interaction
(3) other 'psychological' benefits of exercise (distraction from daily routine and worry, the sense of achievement and social interactions)
(4) increased metabolic rate
(5) increased body temperature
(6) increased heart rate for some prolonged period of time
(7) increased endorphin and adrenaline levels release caused by exercise (Brown et al. 1979; Koch, Johansson, and Arvidsson 1980)

One might reasonably take increased endorphin and adrenaline levels (factor 7) as characteristic for the treatment of depression. Although these features are conceptually distinguishable, the various factors seem to come as a package. It would be difficult, if not impossible, to design a control treatment that contained all the other features of exercise less the increased endorphin and adrenaline levels[58]. That is, the method for constructing placebo controls for drugs, is simply unavailable to exercise-therapy.

In practice, researchers have used other treatment processes as 'placebo' control treatments for exercise, including supervised relaxation (McCann and Holmes 1984) and supervised flexibility (Dunn et al. 2002). The relationship between these control treatments and the experimental intervention they are intended to 'imitate', is very different from the relationship between a pill placebo and the experimental treatment it

[58] In theory one could perhaps inject chemicals that neutralized endorphins and adrenaline into a control group that was taking exercise. This would seem to provide a treatment process that involved all the non-characteristic features of exercise-therapy itself. However these chemicals, even if they existed, would have to be free from further, side effects. Then there would be the problem of a further difference between the control and experimental groups, namely that the experimental group did not receive an injection. To remedy this further problem, the experimental group could be given a sham ('placebo') injection. Then there would be the problem of ensuring that the correct dose of neutralizing chemicals were administered – endorphin and adrenaline levels differ greatly across individuals. All in all, the solution of 'neutralizing' the endorphins and adrenalin levels in the control group, although very interesting, might introduce too many further problems to be a reliable control.

LD/3	low dose of exercise three times per week	7
LD/5	low dose of exercise five times per week	7
PHD/3	public health dose* three times per week	17.5
PHD/5	public health dose five times per week	17.5
'placebo'	supervised stretching for 15-20 minutes three times per week	n/a

*amount of exercise consistent with US public health recommendations

The primary outcome measure was a change in the 17-point Hamilton Rating Scale for Depression ($HRSD_{17}$)[59] score from baseline to 12 weeks. Blind assessors measured the $HRSD_{17}$ scores each week. At the end of the 12 week trial, all groups had some improvement (see table below for summary of results).

6. Table 5.2: Efficacy Analysis after 12 Weeks of Treatment*. (Adapted from Dunn *et al.*'s Table 3 2005).

	Mean Baseline $HRSD_{17}$ (Standard Deviation, SD)	**Mean $HRSD_{17}$ (SD) at 12 weeks**	**% reduction from baseline**	**Significance level**
LD/3	16.2 (4.1)	10.5 (1.2)		n/a
LD/5	16.2 (4.1)	11.9 (1.6)		n/a
LD (overall)			***30%***	***p* = 0.006**
PHD/3	16.2 (4.1)	9.0 (1.0)		n/a
PHD/5	16.2 (4.1)	7.9 (1.3)		n/a
PHD (overall)			***47%***	***p* < 0.001**

[59] The HRSD17 (Hamilton 1967) measures severity of depressive symptoms. It is a questionnaire with 17 questions that ask patients to rate themselves on a 3 or 5 point scale (depending on the question) for symptoms such as insomnia, anxiety, depression, and guilt. A score in the 10-13 point range (out of a possible 69) is taken as evidence that the patient is considered mildly depressed; from 14-17 points, the patient is considered mild to moderately depressed; >17 point, the patient is considered to be severely depressed . The HRSD17 has been widely used to measure depressive symptoms since the 1960's.

Control	16.2 (4.1)	14.0 (4.9)	29%%	$p = 0.02$
Total	**16.2 (4.1)**	**10.0 (0.6)**	**30%**	n/a

* Scores adjusted for age, gender, and baseline score.

The main finding of the study is that a PHD of exercise reduces the scores on the HRSD by an average of 47%. The LD reduces the scores by an average of 30%, which is similar to the 'placebo' treatment, which reduces the scores by an average of 29%. The PHD is similar to the effectiveness of antidepressant medication which reduces the HRSD scores by roughly 42% (Dunn et al. 2005). There is also evidence of a dose-response effect: the LD is not as effective as the more intense PHD. Evidence of a dose response relationship adds to the existing evidence that exercise may have antidepressant properties, including a systematic review (Butler et al. 2003). Given that exercise therapy, unlike many pharmacological therapies for depression, has mostly positive side-effects, this study could have had a huge impact on the default treatment for depression, the problem that most people resist taking exercise notwithstanding.

However the placebo control treatment used in this study might be argued to be illegitimate because it did not permit the trial to remain double masked[60], and therefore did not include any potential effects of expectation of taking real exercise: "participants were unable to be blinded to treatment assignment; therefore, some participants regarded being assigned to the exercise placebo to be unacceptable and immediately dropped out" (Dunn et al. 2005`, p.7). Obviously, the participants in the placebo group deduced that they were in the 'placebo' group. Both participants and investigators supervising the intervention could have had different beliefs about the potency of each intervention. Most likely, the belief that more intense exercise might be more effective was disproportionately present in the treatment groups, while the belief that flexibility was ineffective was disproportionately present in the 'placebo' control group. This

[60] The study has been critiqued on other grounds as well. First, only 5% of the 1664 pre-screened (*prima facie* eligible) patients entered the trial. This suggests that the idea of sweating on a treadmill or bike 3-5 times per week is unappealing to depressed people. Second, there was a 64% drop-out rate in the control group. Once participants discovered they were in the 'placebo' group, they dropped out (usually to seek 'real' treatment). Third, and related to the second, it was impossible to blind both the participants and investigators supervising the interventions. These critiques are irrelevant to the issue of whether the control treatments themselves were legitimate.

means that the expectation effect could have been stronger in the experimental groups and lowest in the 'placebo' group. If participant expectations have effects, and in the treatment of depression it is likely that they do, then the 'placebo' treatment was missing an incidental feature of treatment with the experimental intervention. This is sufficient to question the legitimacy of the placebo control[61].

In addition to missing at least one non-characteristic feature, the control treatments used also had *additional* features that could be effective for depression. There is independent evidence suggesting that supervised flexibility may have an effect on depression. Yoga, for example, which may be little more than flexibility exercises used in trials of exercise, has been shown in at least one trial to reduce the symptoms of depression on the HRSD (Woolery et al. 2004). Then, relaxation (used in other 'placebo' controlled trials of exercise) might also have characteristic effects for depression. Exercise has been linked to the relaxation response in laboratory animals (Choate, Kato, and Mohan 2000). Benson's bestseller, *The Relaxation Response,* first published in 1975 and re-issued in 2000, claims that relaxing for as little as 10 minutes per day can have a positive impact on depression. If it is true that flexibility or

---

[61] It might be argued that placebo controlled trials of exercise can be double masked in spite of the requirement of informed consent. The participants and investigators supervising the treatments could have been informed that participants would be randomized to one of several treatments that *may* lessen depression. These treatments, the participants could be told, include exercise and supervised relaxation. However, this approach would be objectionable on both ethical and methodological grounds. Ethics committees will typically require that participants are informed of the fact that they will enrol in a placebo controlled trial. The suggestion that the participants be randomized to one of several treatments that may lessen depression does not include the information that one of the treatments is (believed to be) a 'placebo' (Emanuel and Miller 2001). It could also, of course, be claimed that supervised relaxation or flexibility could have characteristic features, in which case telling the patients that they would be randomized to one of several treatments that may reduce depressive symptoms would not be misleading the participants. Yet if supervised relaxation or flexibility is believed to have effective characteristic features, then it is no longer legitimate to test exercise against that control treatment in a placebo controlled trial. Recall that the test treatment must demonstrate *superiority* to placebo in order for its characteristic features to be considered effective. If the (alleged) 'placebo' control in fact has characteristic effects of its own, the requirement that the test treatment demonstrate superiority is no longer justified.

relaxation have effects over and above the incidental effects of exercise, then they are illegitimate placebo controls. I am not of course making any claims about the potency of supervised flexibility for depression, but only pointing to the possibility that flexibility may have effects on depression that are greater than the incidental effects of exercise *per se*.

In short, 'surrogate' placebo controls, by definition, do not share much in common with the non-characteristic features of the experimental treatments. If it *could* be shown that these surrogate treatments nonetheless had the same *effects* as the non-characteristic features of the experimental treatment then we could still argue that they were legitimate in spite of the fact that they failed to satisfy the conditions for legitimacy. However precisely because we cannot separate the characteristic from the non-characteristic features of exercise, we cannot separate their effects. We therefore have no way of measuring whether the effects of the surrogate treatment match the effects of the non-characteristic features of the experimental treatment. Therefore we have no way of knowing whether the surrogate placebo control is legitimate. In fact, it would be nothing short of a miracle for a surrogate treatment to be equally (within some appropriate margin) effective as the non-characteristic features of the experimental treatment.

Note that exercise is not the only treatment for which the characteristic/non-characteristic divide is difficult to draw for the practical purpose of designing a legitimate placebo control. In addition to exercise, the characteristic features of many types of psychotherapy (or any form of "talking" therapy) are notoriously difficult to separate from their incidental features[62].

I will now examine, by considering acupuncture trials, the situation where the characteristic/non-characteristic line, although practically possible to draw, is controversial.

---

[62] Placebo controlled trials of psychotherapy, and other forms of 'talking' therapy have been done. However there are two problems with these trials. First, whether the 'placebo' therapy was legitimate is still a subject of debate (Kirsch 2005). I will refrain from delving into this debate, but suffice it to say that the 'placebo' therapies are not as easy to design as pill placebos. Second, the therapists cannot be masked, which introduces a potential confounder on its own and could, as I noted above, subvert the masking of the participants.

5.5. The problem for treatments with disputed therapeutic theories face in meeting 2nd condition: an illustration with acupuncture

Derived from traditional Chinese medicine, acupuncture is a form of treatment for various disorders that involves insertion of fine needles into particular points in the body known as 'Qi' (pronounced 'chee') or 'acupuncture' points. The needles are small and usually penetrate to a depth of about a quarter to three quarters of an inch (5 – 40 millimetres) depending on location. The penetration of the needles into the skin is barely perceptible. The United States Food and Drug Administration has removed its 'experimental' label from acupuncture needles[63], and the National Institute of Health (NIH) of the UK recommends acupuncture as an established treatment for some target disorders, especially those involving pain (NIH 1997).

The treatment factors involved in acupuncture therapy may include the following:

1. Patients and practitioner beliefs about, attitude towards and expectations of needling and acupuncture
2. The acupuncture consultation
3. The physiological effects of (acu)pressure
4. The (putative) effects of (acu)pressure at the correct location
5. General physiological effects of needle insertion (anywhere in the body, not at the 'acupuncture' points indicated by the relevant[64] theory of acupuncture)
6. The (putative) effects of appropriate needle stimulation at the correct location.

Although widely used throughout the world and considered effective for some ailments, acupuncture has rarely been tested in double masked conditions. Most trials are therefore open to the possibility that expectation effects confounded the study. More

---

[63] "The needles used for acupuncture no longer need "investigational use" labeling. FDA recently reclassified them for "general acupuncture use" by qualified practitioners. Acupuncture needles, which are used as part of a centuries-old Chinese healing technique, are medical devices under FDA regulations. Several years ago, the agency reclassified the needles from class III, a category that requires clinical studies, to class II, which means they can be used by licensed, registered or certified acupuncture practitioners. As with other class II devices, the needles are required to have proper labeling, and good manufacturing practices must be followed" (FDA 1996)

[64] Theories of acupuncture are different in China, Japan, and even the United States and Europe.

relevantly for present purposes, it has proven difficult to design a placebo control treatment to be used in acupuncture trials.

Recently, however, several treatments have been developed that allegedly function as 'placebo' or 'sham' acupuncture treatments. One these treatments involves the use of the 'Streitberger Needle'. The Streitberger Needle is a blunt needle that is embedded in a moveable shaft. When the blunt needle is pressed on the skin, the shaft moves and gives the appearance that the needle penetrates the skin. The Streitberger needle in fact, however merely scratches the skin without penetrating. Trials reveal that most patients cannot discern whether they are receiving real or 'Streitberger' acupuncture, and hence that it is 'validated'(Streitberger and Kleinhenz 1998; Kaptchuk et al. 2006). In short, if we ignore the failure of treatment with the Streitberger Needle to keep the dispensing practitioners masked, treatment with the Streitberger Needle seems to perform the functions required for being a legitimate placebo control.

In an interesting study employing treatment with the needle, Kaptchuk and some colleagues at Harvard compared the effects of the Streitberger needle with a placebo pill for the treatment of people with persistent upper extremity pain due to repetitive use. This condition is often called repetitive strain injury, the modern generalized equivalent of "weaver's hand", "sprout picker's thumb", and "scrivener's palsy" (Kaptchuk et al. 2006). In the study of 270 participants, they compared one trial of real acupuncture therapy vs. 'sham' acupuncture (involving the Streitberger needle), and another of amitriptyline therapy vs. placebo pill therapy. The outcomes were measured on a 10-point pain scale where the pain was subjectively reported by the patients. The results indicated

> significantly greater downward slopes per week on the 10 point arm pain scale in the sham device [Streitberger needle] group than in the placebo pill group (-0.33 (-0.40 to 0.26) *v* -0.15 (-0.21 to -0.09), P = 0.0001) and on the symptom severity scale (-0.07 (-0.09 to -0.05) *v* 0.05 (-0.06 to -0.03), P = 0.02). Differences were not significant, however, on the function scale or for grip strength (Kaptchuk et al. 2006).

In short, the Streitberger needle was mildly more effective than the placebo pill. The authors conclude. "A validated sham acupuncture device has a *greater placebo effect* on subjective outcomes than oral placebo pills" (Kaptchuk et al. 2006, emphasis added). The Streitberger needle seems to have successfully controlled for needle insertion, which lies at the heart of the alleged mechanism of action for Chinese

acupuncture. Also, the 'sham' acupuncture seems to have been more effective than the placebo pill, at least for certain outcome measures.

However, an important distinction must be made between treatments that control for all but one (or two) non-characteristic features, and treatments that control for all non-characteristic features. As a control for whether needle insertion at the correct acupuncture points (defined by the appropriate therapeutic theory), treatment with the Streitberger needle is a good candidate. The only potential objection to this view would be that the dispensing practitioners were not masked. However, taking this sham treatment as a legitimate *placebo* controls implies that we hold the only characteristic feature of acupuncture to be needle insertion.

It has been argued that needle stimulation at the correct Qi points might not be the only characteristic features of treatment with acupuncture. It could be argued, for example, that acupressure, achieved by the Streitberger needle, should be classified as characteristic. If so, then the Streitberger needle is not a legitimate placebo for acupuncture. The Streitberger needle applies pressure at acupuncture points. There is independent evidence that acupressure is effective for pain relief (Lee et al. 2008; Agarwal et al. 2005). I will not discuss this literature here, but only note that it is not *prima facie* obvious that acupressure should be ruled out as characteristic. If it should be characteristic, the Streitberger Needle is not a legitimate placebo, but only a good control for the putative effects of needle insertion at the acupuncture points. With acupuncture therapy, unlike drug treatments, the selection of characteristic features is not straightforward.

The problem with designing legitimate placebo control treatments for acupuncture trials was also present in another study where a different sham needle was employed. The German Acupuncture Trials (GERAC) investigators compared real acupuncture treatment, 'sham' acupuncture treatment, and conventional therapy. Real acupuncture consisted of bi-weekly treatment that included a 30-minute consultation and insertion of 14-20 acupuncture needles to the 'correct' depth, i.e. 5-40 mm depending on location. 'Sham' acupuncture was identical to real acupuncture apart from two features. First, the needles penetrated only to a depth of 1-3 milimetres rather than the standard 5-40 milimetres (Haake et al. 2007). The justification for this was:

> Induction of Qi (the sensation felt when an acupuncturist reaches the level of Qi [numb radiating sensation indicative of effective needling] in the body) was elicited by manual stimulation (Haake et al. 2007)

This means that the 'sham' acupuncture did not induce the dull radiating sensation. Failure to induce the dull radiating sensation was taken as evidence that the Qi had not been 'activated'. Stimulation of the Qi points for real acupuncture was assured by deeper penetration as well as manual stimulation (moving the needle around manually). Second, the 'sham' acupuncture avoided all known acupuncture points. In short, the 'sham' acupuncture did not contain what many would argue is the distinguishing feature of acupuncture, namely stimulation of Qi points. Although they do not provide details, the authors claim that the sham needle was able to keep the participants masked.

However, like with treatment involving the Streitberger needle, the distinction between a control treatment that serves to isolate the potential effects of a single factor, and a control treatment that serves to isolate the potential effects of all characteristic features must be made. It could been argued, for example, that the therapy with the sham needle used in the GERAC trial was not a 'sham' treatment, but a in effect equivalent to a different form of acupuncture practised in Japan where the needle does not penetrate the skin as deeply as other forms of acupuncture. If the effectiveness of acupuncture does not depend as much on the depth of the insertion, then the evidence for Japanese acupuncture's effectiveness (Birch and Jamison 1998) could be viewed as support for this view.

However, Japanese acupuncture inserts the needles at certain specific points. Unless the GERAC sham acupuncture mysteriously hit these Japanese points while trying to avoid the Qi points (if indeed they are different), then the GERAC sham acupuncture could not be viewed as similar to Japanese acupuncture.

Nonetheless, even if the GERAC sham acupuncture effectively avoided all Qi points, there is also independent evidence that stimulation of the skin by any needles, no matter where they are inserted, affect the central nervous system in a way that reduces pain by activating receptors *anywhere* in the skin (Lundeberg and Lund 2007). I will not evaluate the claims that Japanese acupuncture, or 'needling' should be classified as characteristic in any detail here. It is sufficient to note that it is unclear whether the (implied) therapeutic theory which lead to the design of the GERAC sham acupuncture is mistaken if the general effects of needling and/or shallow needle insertion are characteristic.

To make matters more complex, it is sometimes claimed that the acupuncture *consultation* should be considered characteristic:

> Although some aspects of talking and being listened to are incidental (such as focused attention and empathy), other aspects are characteristic of acupuncture and its underlying theory. For example, the way that a history is taken at the initial consultation indicates to patients that everything about them [i.e. more than what would be considered relevant in a conventional consultation] is relevant to the diagnosis and treatment plan. During subsequent treatment sessions needle insertion and healthcare advice are often varied to take into account any new concerns, whether physical, emotional, or social. This type of talking and listening may result in an increasingly participative interaction in which the whole burden of illness can be [or at least could be] shared and partially relieved (Paterson and Dieppe 2005`, p.1203).

Then immediately below,

> This difference [between the acupuncture and conventional consultation] is not a function of the individual practitioner, ... , but is a function of the different theoretical models underlying biomedicine and acupuncture (Paterson and Dieppe 2005`, p.1203).

The claim that the acupuncture consultation is a characteristic feature might deserve to be taken seriously. If the acupuncture consultation deserves to be classified as characteristic, then a legitimate control for acupuncture could not include the acupuncture consultation.

As a comment on the potential objection that consultations cannot be classified as characteristic, the temptation to rule out all consultations as characteristic must be avoided. Recall Grünbaum's insistence that nothing is a placebo (or, by implication, an incidental feature) *simpliciter*. Cognitive Behaviour Therapy (CBT), for example, consists of nothing more than consultations, yet we do not want to rule them out as nonplacebos without any investigation.

In sum, there is an important distinction between a treatment that successfully serves to control for all but a non-characteristic single factor, and a treatment that successfully serves to control for all non-characteristic factors. In the case of drug trials, this distinction does not usually need to be made explicit because there is usually one (or at most a couple) of easily identifiable, and physically separable, characteristic features. But for more complex interventions such as treatment with acupuncture, identifying which factors from among the many candidates, should be classified as characteristic, is more difficult.

5.6. The Confusion Caused by Heterogeneous 'Placebo Controls'

Even if all placebo control treatments were legitimate, placebo controls would still be problematic because they lead to confusion when it comes to indirect comparisons of

competing interventions for the same ailment. I will illustrate this confusion by considering the GERAC trial in further detail.

The conventional therapy that was compared with real and 'sham' acupuncture in the GERAC trial was performed according to German guidelines and included bi-weekly consultations with physicians or physiotherapists who administered exercise or physiotherapy. Conventional therapy was also "supported by nonsteroidal anti-inflammatory drugs [NSAIDs] or pain medication up to maximum daily dose" (Haake et al. 2007, p. 1893).

> The primary outcome measure was
>
> treatment response after 6 months after randomization, defined as 33% improvement or better on 3 pain-related items on the Von Korff Chronic Pain Grade Scale or 12% improvement or better on back-specific functional statues measured by the Hanover Functional Ability Questionnaire (Haake et al. 2007, p. 1894).

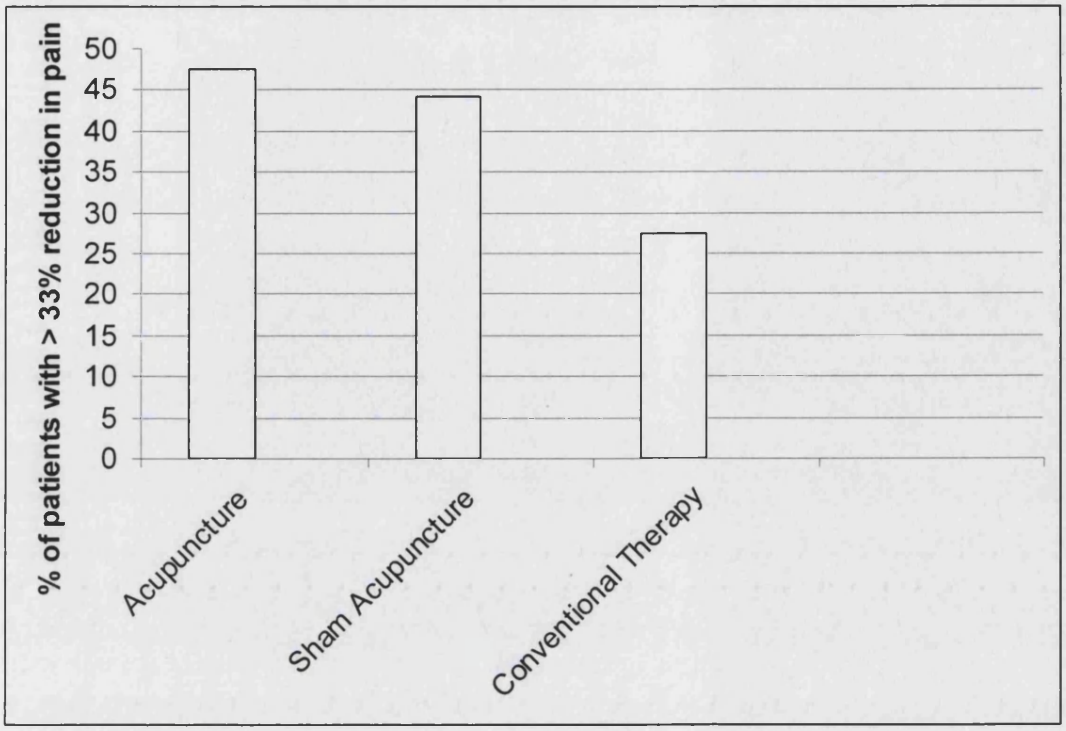
In brief, outcome measures were subjective measures of pain, i.e. '10' means excruciating pain and 0 means no pain. The response rate after 6 months, which was assessed by masked investigators, was

> 47.6% in the verum [real] acupuncture group, 44.2% in the sham acupuncture group, and 27.4% in the conventional therapy group. Differences among the groups were as follows: verum vs. sham, 3.4% (95% confidence interval -3.7% - 10.3%; P=.39); verum vs. conventional therapy, 20.2% (95% confidence interval 13.4% - 26.7%; $P < .001$); sham vs. conventional therapy, 16.8% (95% confidence interval 10.1% - 23.4%; P<.001).

In short, both real and sham acupuncture therapy were significantly more effective than conventional therapy, although the difference between real and sham acupuncture was not significant. The results can be summarized pictorially in the following chart:

7. Chart 5.2: Acupuncture, 'sham' acupuncture, and conventional therapy[65]

[65] Although I arrived at the confusion that could result from different placebo effects for treatments for the same target disorder independently, I borrowed the term 'paradox of effectiveness' used to describe the same thing from Wallach et al (Walach et al. 2006)



The authors conclude that the sham acupuncture "may be a kind of superplacebo effect produced by placebo and all non-specific factors working together" (Haake et al. 2007, p. 1893).

But even if the GERAC sham acupuncture was a legitimate placebo, calling treatment with the 'sham' needle a 'placebo' treatment is confusing. On the one hand acupuncture was no more effective than acupuncture 'placebo', which suggests that acupuncture is a 'mere' placebo and perhaps not worth suffering the costs or side effects. Meanwhile there is evidence that conventional therapy, especially where it involved NSAIDs, is more effective than a (different) placebo[66]. Conventional therapy therefore seems to have characteristic features that have effects and costs worth suffering. Yet even acupuncture 'placebo' was more effective than conventional therapy[67]. Since treatments referred to as 'placebo controls' can be, and often are, very

---

[66] Although comparison between conventional therapy and placebo was not made directly in the GERAC trial there is independent evidence that, at least as far as treatment with NSAIDs are concerned, that the conventional therapy used in the trial is more effective than 'placebo' (Sprott et al. 2006; Coats et al. 2004).

[67] Although comparison between conventional therapy and placebo was not made directly in the GERAC trial there is independent evidence that, at least as far as treatment with NSAIDs are

different and have different effects, they do not provide a consistent baseline for use in making comparisons about which treatments, from among available alternatives, are most effective.

The problem with the confusing comparisons caused by lumping heterogeneous treatment processes into a single category, i.e. 'placebo' controls, can be generalized. One experimental treatment process, call it A, could be no more effective than legitimate placebo treatment process, call it A'. Meanwhile, another treatment process B could be much more effective than B'. Patients, practitioners, and policy makers could be forgiven for concluding, on the basis of separate trials (A vs. A', and B vs. B'), that B was more effective than A. Yet, as the example of the GERAC trial suggests, such a conclusion could be radically false[68], and leads to what is sometimes called the 'double positive paradox', whereby both experimental and 'placebo' treatments have similar effectiveness, both are more effective than another experimental treatment which itself is more effective than a different placebo control.

In short, even trials employing legitimate placebo control treatments cause confusion when it comes to cross-trial comparisons - even legitimate placebo controls are problematic from this pragmatic point of view. I will now proceed to discuss a solution to the problems with placebo controls.

5.7. A Solution to the Problems with Illegitimate Placebo Controls and the Double Positive Paradox: Double Abandonment of 'Placebo Controls'

The two problems with 'placebo' controls can be summarized as follows:

(1) Many treatments described as 'placebo' controls are illegitimate (especially, but not exclusively for non-drug 'placebos') and lead to inaccurate estimates of characteristic effects. The practical difficulties in designing legitimate 'placebo' controls, especially but not only for non-drug treatments means that illegitimate 'placebo' control are not likely to disappear.

---

concerned, that the conventional therapy used in the trial is more effective than 'placebo' (Sprott et al. 2006; Coats et al. 2004).

[68] The GERAC trial does not measure this directly but there is ample independent evidence that NSAIDs, which formed part of the conventional therapy, are more effective than treatment with NSAID placebos for lower back pain (Sprott et al. 2006; Coats et al. 2004).

(2) Different 'placebo' controls, even if legitimate, can have very different effects, which creates confusion when making a comparison between two separate placebo controlled trials.

Rather than recommending ways to face these problems head-on, which the proceeding discussion suggests would be difficult at best, and a blind alley at worst, I will suggest that where 'placebo' controls are likely to be illegitimate, that we test our treatments using other methods, such as dose-response trials or 'active' controlled trials. 'Placebo' controls are likely to be illegitimate where (a) the experimental treatments has side effects that are difficult or unethical to imitate, and (b) where the characteristic / non-characteristic line is difficult to draw.

To avoid inaccurate estimates of characteristic effectiveness caused by illegitimate 'placebo' controls, the control treatments should be described in detail and not referred to as 'placebos'. More specifically, the description of the control treatment should include detailed information about the treatment process, and a specification of features the control treatment is intended to 'control for'. For example, instead of describing the control therapy in a Prozac as a 'placebo control', investigators should provide a more precise description such as 'a treatment involving pill that appears the same as a Prozac pill, but where fluoxetine is replaced by lactose, which is delivered in the same way as real Prozac, by masked investigators to masked participants, whose purpose is to control for all but the effects of fluoxetine ...'. The treatment, once adequately described, should subsequently be referred to as the control treatment rather than the 'placebo'. Although in the case of drug trials this may seem like overkill, the case cited by Golomb shows that it could be necessary more often than we think. In cases of non-drug treatments such as exercise therapy or acupuncture therapy, rather than the terms 'sham exercise' or 'sham acupuncture', which are highly ambiguous, a description such as 'core flexibility exercises three times per week supervised by unmasked, qualified physiotherapists' would help prevent confusion.

Abandoning the general term 'placebo' in favour of more detailed descriptions seems to solve the first problem with 'placebo' controls. Treatments that are not placebos cannot be legitimate (or illegitimate) placebo controls. In addition, quite obviously, a more precise description of the control condition is more accurate. Further, scientific experiments should be replicable; accurate descriptions of the control treatment trials will improve replicability.

Yet, although a control treatment not described as a placebo control cannot be an *illegitimate* placebo control, the underlying problem could well persist. The new control treatment processes, after all, function as placebo controls – their aim is to control for all non-characteristic features of the experimental treatment process. In practice many of the problems creating legitimate 'placebo' controls are likely to plague the better-described controls that function as placebo controls. For instance, take the control treatment involving a pill that appears the same as a Prozac pill, but where fluoxetine is replaced by lactose, which is delivered in the same way as real Prozac, by masked investigators to masked participants, whose purpose is to control for all but the effects of fluoxetine ...'. This control treatment might not successfully be delivered in masked conditions. If not, then the better-described control treatments can be illegitimate in the same way 'placebo' controls were.

Worse, problems with the double positive paradox could well persist even in trials where the control treatment is adequately described. If experimental intervention X failed to demonstrate superiority to control treatment X' (a control treatment functioning as a placebo control), while experimental treatment Y demonstrated superiority to control treatment Y' (a control treatment functioning as a placebo control) many would conclude that Y was more effective than X. As we saw above, such a conclusion could be radically wrong. Hence abandoning the term 'placebo' in favour of a more precise description of the control treatment, although a definite improvement over the current state of affairs, is insufficient to overcome the two problems with 'placebo' controls.

To truly avoid the problems, 'placebo' controls, or any control treatment whose purpose is to control for all non-characteristic features, should be abandoned altogether in favour of active controls, at least in cases where placebo controls are likely to deceive.

The use of active controls circumvents both problems with placebo controls. In an active controlled trial, the experimental intervention must prove that it is as effective as, or more effective than, one or more standard control treatments. The conclusions drawn from an active controlled trial are not about the effects of the characteristic features, but about the relative effectiveness of the overall experimental treatment process compared with the overall effectiveness of the control treatment process. This avoids the first problem because no estimate of the characteristic effectiveness of the

experimental treatment is provided. The second problem is solved because a direct comparison between two or more non-placebos is provided.

Active controlled trials, however, have alleged methodological problems of their own. The most important is that, if one treatment is as or more effective than a control treatment that is believed to be effective, the interpretation is ambiguous. Both treatments could be effective or both treatments could be ineffective. For instance, if I found that lobotomies were as effective as bloodletting for the treatment of depression, it would be incorrect to conclude that both interventions were effective. They could both be ineffective, or worse, harmful. Then, if we compare interventions that are very different, such as, say, exercise and pills, these active controlled trials will be difficult to perform in double masked conditions.

With the conceptual work of defining placebos and placebo controls complete, I am well placed to employ 'scientific common sense' to evaluate the arguments that 'placebo' controls are superior to 'active' controls. Since, however, a discussion of the methodological role of double masking in many ways informs the debate about the alleged superiority of placebo over 'active' controls, I will dedicate the next chapter to an investigation of double masking.

6. Chapter Six. Double-Blinding: The Benefits and Risks of Being in the Dark

The patient, treated on the fashionable theory, sometimes gets well in spite of the medicine. The medicine therefore restored him, and the young doctor received new courage to proceed in his bold experiments on the lives of his fellow creatures"
- Thomas Jefferson (Jefferson and Irwin 1975)

many investigators and readers delineate a randomized trial as high quality if it is "double-blind," as if double-blinding is the sine qua non of a randomized controlled trial. ... A randomized trial, however, can be methodologically sound ... and not be double-blind or, conversely, double-blind and not methodologically sound.
- (Schulz, Chalmers, and Altman 2002)

6.1. The Problems with Double Masking as a Requirement for Clinical Trial Validity

Being 'double blind' or 'double masked', where neither the participants nor the investigators are aware of who gets the experimental treatment, is almost universally trumpeted as being a virtue of medical experiments. Sir Austen Bradford Hill, for example, states:

> By the so-called double-blind procedure, when neither patient nor doctor knows the nature of the treatment give to an individual case, it is hoped that unbiased subjective judgements of the course of the illness can be obtained (Hill and Hill 1991, p. 214).

The official EBM text states:

> Blinding is necessary to avoid patients' reporting of symptoms or their adherence to treatment being affected by hunches about whether the treatment is effective. Similarly, blinding prevents the report or interpretation of symptoms from being affected by the clinician's or outcomes assessor's suspicions about the effectiveness of the study intervention (Straus, Richardson, and Haynes 2005, p.122).

Armitage, Berry, and Lindgren state:

> In such a trial [i.e. a placebo controlled trial] the mere knowledge that an additional intervention is made for one group only may produce an apparent benefit, irrespective of the intrinsic merits of that intervention. The principle of *masking* the identity of a treatment may be extended to trials in which two or more potentially active treatments are compared. The main purpose here is to ensure that the measurement of the response variable is not affected by knowledge of the specific treatment administered (Armitage, Berry, and Matthews 2002).

Although as we shall soon see the term 'double masked' is often used ambiguously, support for double masking can be found in virtually any account of clinical trials textbook (Greenhalgh 2006; FDA 2005; Bland 2000; Jadad 1998). This is because failure to double mask the study successfully introduces the possibility that participant or investigator beliefs or expectations confound the study. As a consequence, failure to keep trials double masked is regarded as a relative vice.

Although double masking certainly adds value in many cases, it must be recalled that there are problems with viewing double masking as a universal virtue. For one, it leads to the "Phillip's paradox" (Ney, Collins, and Spensor 1986), that very effective experimental treatments will not be testable in double masked trials – the dramatic effectiveness of the treatment will make it identifiable by both trial participants as well as dispensing physicians. For instance, imagine a new drug for the common cold was invented that the investigators were intending to test against an (initially) observationally-indistinguishable placebo. If the new pill made even the most severe symptoms of the cold disappear within seconds of swallowing it, then most participants and investigators would correctly guess which treatment was the wonder-drug and which was the placebo. One may object that the "Phillip's paradox" is hardly problematic and merely re-states the obvious point that very effective treatments do not require testing in highly controlled (i.e. double masked) conditions. Nonetheless, the paradox indicates that according to the *unqualified* view that double masked trials are superior to 'open' (unmasked) trials, our best treatments cannot be supported by the best evidence.

A further problem is that many treatments, ranging from most surgical techniques to exercise, cannot be tested in double masked conditions. For these treatments double masking is an impossible standard. Although we may in the end have to bite the bullet and admit that these treatments are simply unlucky and cannot be supported by 'best evidence', it is surely worthwhile first investigating the alleged virtue of double masking very carefully. Perhaps because the terms 'double masking', 'placebo controls' and even 'randomized trial' are often spoken of as if they were necessarily connected, the methodological virtues of double masking have seldom, if ever, been systematically examined in isolation.

In this chapter I will evaluate the role of double masking from the fundamental view that good evidence rules out plausible rival hypotheses. To anticipate, I will argue that when investigated this way, it is clear that the methodological value of double

masking is far more limited than is usually admitted. After a few clarificatory remarks about the meaning of double masking, I outline the rationale for the view that double masking increases the internal validity of a study. In short, it is thought that two potential confounders, participant and investigator expectations, can be eliminated by successful double masking. If the investigator is aware that a particular participant is in the experimental arm of the trial they may lavish more attention on them[69]. This increased attention could have therapeutic benefits for certain ailments. Similarly, if the participant believes she is receiving the best treatment (as opposed to the placebo), then her knowledge that she is in the experimental arm could lead her not only to report better outcomes, but to experience greater beneficial effects. I then point out that these two *potential* confounders are sometimes not *actual* confounders. Then, I contend that there are severe practical limits to the potential success of attempts to keep trials double masked. If so, then there is little value in being *described* as double masked. Finally, double-masking could impair external validity, since it contributes to making the trial importantly different from routine clinical practice. In conclusion, double masking, although it potentially increases the internal validity of a study, does not always do so; further, since double masking may not be possible, we may be better off seeking other ways to control for the potentially confounding effects of expectations.

6.2. The Many Faces of Double Masking: Clarifying the Terminology

The term 'double masked' is used in several different ways to describe the masking of various groups involved in a clinical trial. It is therefore necessary to make some clarifying remarks about how I will use the term.

First, however, I will defend my use of the term 'masked' instead of the more common 'blind'. The term 'blind is ambiguous in trials of blind people, and it is especially abhorred by researchers of eye disease (Bland 2000, p. 19). Second, 'masking' someone implies that the concealment procedure could be imperfect. As I will argue later, the process of concealing knowledge to study groups is less successful than most of us believe, and indeed may be inherently difficult to achieve. Third, the

[69] The experimenter could also encourage them to remain in the trial. Intention-to-treat analysis means that drop-outs in the experimental arm tend to reduce the apparent effectiveness of the experimental treatment.

term 'masking' is more in line with the historical meaning. Early trials that concealed the nature of the treatments from participants literally used masks (Kaptchuk 1998).

Masking is the act of concealing the nature of the intervention from one of the groups involved in the study. For example, in a single masked randomized trial of vitamin C versus placebo as a cure for the common cold, the participants in the trial could be prevented from knowing whether they were taking the placebo or real vitamin C.

Six groups involved in a trial that are sometimes masked, namely:

1. **Participants**
2. **Intervention dispensers (henceforth "dispensers"):** This group administers the intervention, whether it is medical or not. Doctors and nurses performing a surgical intervention are dispensers. Psychiatrists and exercise trainers might also be dispensers.
3. **Data collectors**: The individuals responsible for collecting the data for the study outcomes. This could include taking a blood pressure measurement, reading an X-ray, administering a questionnaire, or "recording symptoms potentially compatible with a transient ischemic attack" (Montori et al. 2002, appendix).
4. **Outcome evaluators:** This group decides if the participant has suffered (or enjoyed) the outcome of interest. For example, they decide whether a participant has died, or whether a patient has high blood pressure. The step of evaluating the outcomes often goes together with collecting data. For example, in a trial of an antidepressant drug, the data collector might administer the HRSD and then analyze the data. However, they are separate in many cases, and could be separated in many others, so for purposes of analysis I will follow Devereaux (2001) and use a separate category for outcome evaluators.
5. **Data analysts**: These are the statistical analysts who make decisions about the type of statistical tests to perform and then perform the tests.
6. **Personnel writing the paper**: This group is very rarely masked. The personnel writing the paper are those who might write alternative versions of the manuscript before the trial is unmasked. In the simple case where the trial has one experimental and one control group – call them A and B respectively, one paper is written as if A is the experimental and another is written as if B is the experimental intervention. There is some evidence that whether the experimental intervention was found to be effective or not can influence how the manuscript

is written. This is usually but not necessarily done with the collaboration of the data analysts.

In a study of physicians, 5 major journals and several textbooks, Devereaux found that the term double masked is used in over a dozen ways (Devereaux et al. 2001). About a third defined "double masking" as masking of the participants and dispensers. The remaining definitions included various combinations of 2, 3, and 4 masked groups. Because of this ambiguity, both the CONSORT Statement (Moher, Schulz, and Altman 2001) and Jadad (Jadad 1998) recommend identifying the particular groups that have been masked rather than using the terms 'single masked', 'double masked', or 'triple masked', etc.

Although specifying exactly which groups have been masked is always useful, there are good reasons to reserve the term double masked for trials that mask the participants and the dispensers. For one, only the knowledge or beliefs of these two groups can be considered as features of the treatment process *per se*: the participant's or dispenser's belief that a particular patient is being treated with the experimental treatment may be a feature of the treatment process that could have direct effects on the target disorder, while knowledge on the part of the other groups are not. My belief that I am getting the 'real' as opposed to placebo, or older treatment, can directly affect the target disorder – or at any rate how I feel about my recovery. Similarly, the belief of the dispenser that a particular participant is getting the 'real' or best treatment can translate into both stronger participant belief as well as different treatment. Precisely how the participants and dispensers can be part of the treatment will become clearer in the next section. Knowledge of the data collectors, and certainly the knowledge of the other groups, does not affect the treatment process in any straightforward manner. Moreover, the only two groups that cannot be masked in an *observational study* are the participants and dispensers. Doctors and patients in routine clinical practice are supposed to know what treatment is being administered. It is, on the contrary, unproblematic, even methodologically desirable, to mask the data collectors, outcome evaluators, data analysts, and even manuscript writers in observational studies. Relatedly, the participants and dispensers are most difficult to mask in trials of treatments that are not easily imitated by legitimate placebo controls. For example, in a trial of exercise versus "placebo" (say supervised flexibility) for depression, although it is problematic to mask participants and dispensers, there is no reason why the other groups cannot be masked.

Hence I will use the term double masked to refer to trials where the participants and dispensers are masked.

Reserving the term double masked for trials where the participants and dispensers are masked emphatically does not mean that masking the other groups is unimportant. I will not get into a debate about the merits of masking these other groups[70]. Suffice it to note that masking the other groups may well rule out confounders and that it is therefore important to attempt to achieve masking of these groups. More relevantly, any arguments I present about the limited value of double masking do not bear on the importance of masking the other groups.

Concealed allocation is sometimes confused with masking. In fact, concealed allocation "occurs when the person who is enrolling a participant into a clinical trial is unaware whether the next participant to be enrolled will be allocated to the intervention or control group" (Straus, Richardson, and Haynes 2005, p.279). For example, in a RCT with 10 participants, some process could allocate the 2nd, 3rd, 6th, 8th, and 9th participant to group A and the 1st, 4th, 5th, 7th, and 10th participant to group B, but not indicate whether A was the experimental or control intervention. Subsequently (post-allocation), the same or different investigators as well as the participants could be made aware of which was the experimental intervention and which was the control. Note further that, as in the above case, the allocation need not be random in order for it to be concealed. Although random allocation might make concealment easier, and concealed allocation might make masking easier, the concepts are distinct[71].

---

[70] For example, assessing outcomes on X-Rays or taking blood pressure can be influenced by beliefs of the outcome assessor (Sackett 1991). In conversation, many researchers admit that they hire several data analysts and choose the results of the one they like best. Masking the statisticians and manuscript writers would prevent this. Then, if the results were not what the authors expect, the way they write the article could colour the data. In an example I will discuss in detail later, it could be argued that Hróbjartsson and Peter Gøtzsche colour their conclusions of the magnitude of the placebo effect (see below). The potential confounding of these unmasked groups remains to be studied in any detail. "The frequency and magnitude of ascertainment bias introduced after data collection have not been studied at all" (Jadad 1998, p.55).

[71] Concealed allocation plays a role in preventing selection bias, while masking the participants and dispensers helps prevent performance bias; masking the other groups attempts to rule out assessment bias (see chapter 1).

To sum up, although the term double masked is currently used in conflicting ways, there are good reasons to reserve the term double masked for trials whose participants and dispensers are masked. Why then should double masking as thus characterised be considered important?

6.3. Participant Expectation and Pygmalion Effects as Confounders

In this section, I will explain why it is commonly held that beliefs of participants and dispensers that they are receiving/dispensing the experimental intervention can confound a study. First, however, I will clear up a confusion between what are commonly called 'placebo effects' and participant expectations.

6.3.1. Mistaken Estimates of the Placebo Effect

If participants and dispensers have beliefs about the effectiveness of the experimental treatment then since these are part of the placebo effect, it is likely that there will be placebo effects. The placebo effect is typically estimated by measuring the effects of placebo treatments. Although placebos have been used knowingly and unknowingly by physicians for at least several centuries, Henry K. Beecher was the first person to attempt to quantify placebo effects in his work which was immortalized in his famous article "The Powerful Placebo" (1955). Beecher's article is still the most cited work about placebos. His method was to measure the effectiveness of the placebo control in the control group, and take that to be the effects of the placebo control treatment. In his study he reviewed 15 studies of treatments for post-operative pain,

---

Although randomization, concealed allocation, and masking are conceptually distinct, they often confused in the medical community. The prominent medical statistician Stephen Senn, for example, states: "I am interested in showing how randomization is required to support blinding, but it is worth noting that the converse may also be the case" (Senn 1994, p. 221). Senn goes on to cite a study that suggested that "treatment allocation was biased unless the trials were randomized and blind" (ibid, p. 221). As is quite clear, however, there is no sense in which concealed allocation is required by subsequent blinding. It may be the case, of course, that trials which tend to be unblind also tend not to used concealed or random allocation and generally to be of poorer quality.But this, if true, is a merely contingent matter and not necessitated by logical connections between these notions.

cough, pain from angina pectoris, headache, seasickness, and anxiety. The studies had a total of 1082 participants, and found that overall, 35% (± 2.2%) of the patients' symptoms were relieved by placebo alone (Beecher 1955, p. 1604).

However, it is a mistake to attribute any change in the placebo control group to the placebo control treatment. Beecher has been criticized for failing to consider and rule out explanations for the improvement in the placebo control group other than expectation or placebo effects (Kienle and Kiene 1997). In particular Beecher failed to consider the natural history of the disease, which may of course include spontaneous improvement[72].

Many ailments, including the common cold and post-operative pain usually go away quite quickly without any treatment at all. More generally, the natural history of the disease, and spontaneous remission, are all potential causes of apparent recovery that have nothing whatsoever to do with placebo effects. Many diseases vary spontaneously with time – the "natural history of the disease". Spontaneous remission is a common reason for symptoms of an ailment to improve that has nothing to do with the placebo effect.

In order to estimate the effect of the placebo treatment, the effects of these other potential causes of recovery in the placebo control group must be taken into account, something Beecher failed to do. When, for instance, 35% of patients with mild common colds felt better within 6 days (2 days after the onset of placebo administration), Beecher concluded that the effects were due to the placebo administration. Beecher "did not consider that many patients with a mild common cold improve spontaneously within 6 days" (Kienle and Kiene 1997, p.1312). In another similar case,

> Beecher referred to patients with diseases such as ulcer, migraine, muscle tension, or headache who suffered from anxiety and tension and were treated for eight 2-week periods alternatively with mephenesin and placebo. Beecher claimed a placebo [expectancy] effect of 30% since "roughly" 20-30% of the patients improved" (Kienle and Kiene 1997, p.1313).

In these examples, Beecher makes the error of failing to take into account spontaneous remission – it is wrong to conclude from the fact that 20-30% of patients improved, that the improvement was due to the effect of expectation.

---

[72] 'Observer bias' which is bias of investigators observing the outcomes, is often included as a potential explanation for placebo effects. Because the genealogy of observer bias is distinct, I discuss it separately below.

Kienle and Kienle examined all the studies upon which Beecher based his estimate of the placebo effect and concluded that "none of the original trials cited by Beecher gave grounds to assume the existence of placebo effects" (Kienle and Kiene 1997, p.1316). Kienle and Kienle conclude that "the extent and frequency of placebo effects as published in most of the literature are gross exaggerations" (Kienle and Kiene 1997, p.1316). I will not repeat Kienle and Kienle's detailed study of each of the trials considered by Beecher. Suffice it to say that Beecher failed to rule out natural history of the disease as a plausible hypotheses for effects measured in the placebo control group and that his commonly cited estimate of the magnitude of the placebo effect may thus have been exaggerated.

In defense of Beecher, it might be argued that he had more evidence for placebo effects than Kienle and Kienle allow, and the scope of his conclusion was not as far reaching as is sometimes assumed. Supporting the hypothesis that placebos have effects, he cites (qualitative) evidence of a dose-response effect in different placebos - this evidence is not called into question by any of the points that Kienle and Kienle make. Also, he limited the scope of his argument to treatments with *subjective responses*. "It is evident that placebos have a high degree of therapeutic effectiveness in treating subjective responses" (Beecher 1955, p. 1604).

6.3.2. Participant belief

A participant's belief that she is being treated with an effective drug could, at least in theory, translate into effects for the outcome of interest. For example, if I believe I am being given the latest and best treatment for the common cold, I may well recover more quickly than had I not taken the latest treatment, or I may well report that I have recovered more quickly and this is all that is at issue when the outcome is subjective.

I will call the effects of knowledge that one is being treated with something one believes at least may be effective 'belief effects'[73]. To measure the effect of participant belief, we need a trial where one group of participants knows they are receiving the

[73] These are also commonly called expectation effects'. However, there is a debate about whether it is the expectation, conditioning, or meaning of the treatment that are responsible for the effects (Moerman and Jonas 2002; Kirsch 2004; Benedetti et al. 2004). I will not delve into this debate here. The salient feature of all these possible mechanisms is that the participant must have some kind of awareness that she is being treated.

intervention, while another group does not believe they receive the intervention. Recent studies of analgesics employed such a design. Using a truncated version of the balanced placebo design[74], Benedetti and a team of researchers at the University of Turin treated patients "overtly" and "covertly" for postoperative pain, Parkinson's and anxiety. I will focus on the case of postoperative pain.

In a study of pain (Benedetti et al. 2004) Benedetti's team used four common painkillers - buprenorphine, tramadol, ketorolac, and metamizol - on a total of 278 patients who had undergone thoracic surgery for different pathological conditions. The postoperative patients were taken to have provided their 'informed consent' when they were "told that they could receive either a painkiller or nothing depending on their postoperative state and that they will not necessarily be informed when any analgesic treatment will be started" and they agreed to be in the study. "In this way, patients [did] not know if or when the treatment [was] given." (Benedetti et al. 2004, p. 680).

The patients were then, of course unbeknownst to them, randomized into 'overt' and 'covert' groups with sex, age, weight, and pain baseline-balanced. The 'overt' group was treated by doctors who "gave the open drug at the bedside, telling the patient that the injection was a powerful analgesic and that the pain was going to subside in a few minutes" (Benedetti et al. 2004, p. 681). Then, one dose of analgesic[75] was administered every 15 minutes until a 50% reduction of pain (from baseline) was achieved for each patient. The 'covert' group, on the other hand had the analgesic delivered by a pre-programmed infusion machine (already attached to the patient) without any doctor or nurse in the room. The pain reduction for both sets of patients was measured every 15 minutes on a 10-point subjective pain scale where 0 = no pain and 10 = unbearable pain.

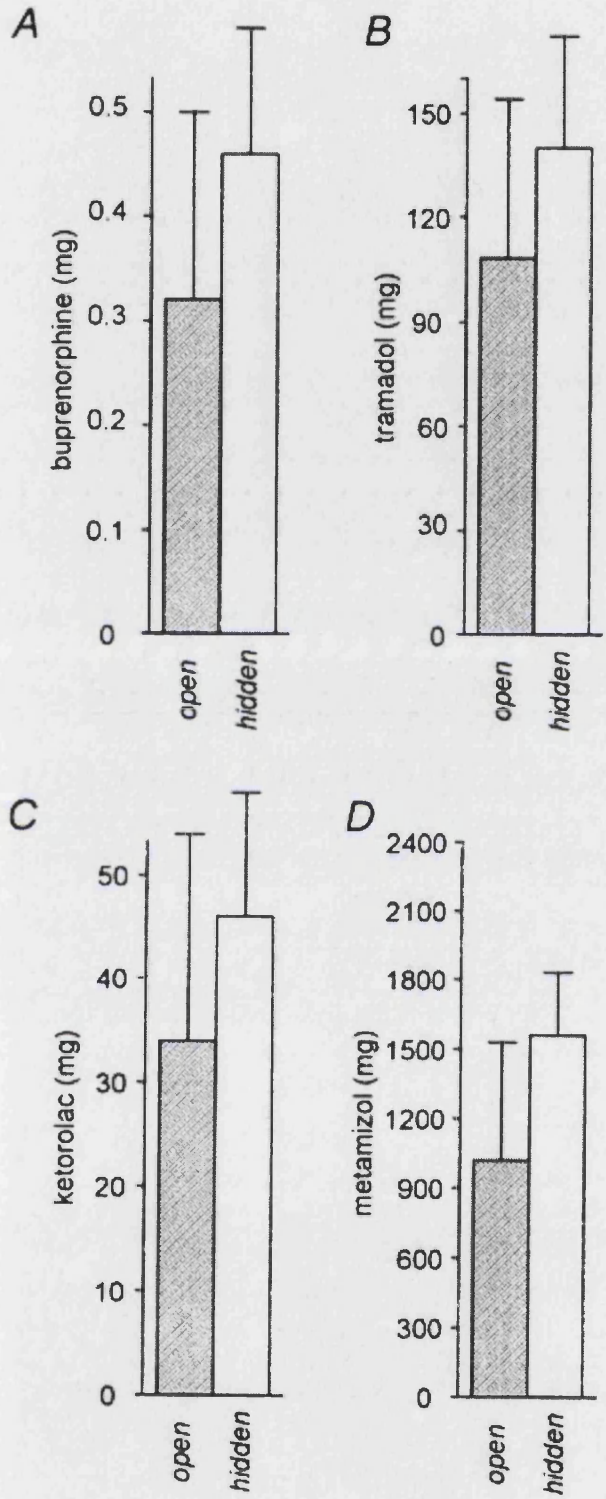
The results were that over 30% more analgesic was required by the patients who were treated covertly (p-values ranging from 0.02 – 0.007 depending on drug). See the figure below for details.

---

[74] See chapter 7 for a full description of the balanced placebo design.

[75] Doses of different analgesics were standardized according to a method described earlier (Benedetti et al. 1998).

8. Figure 6.1: The amount of analgesic required to reduce pain by 50% for buprenorphine (A), tramadol (B), ketorolac (C), and metamizol (D). From (Amanzio et al. 2001, p. 209).



Benedetti's study has been criticized on the grounds that the patients in the covertly treated group may have detected when they were getting treated in spite of the attempt that it was done 'covertly'. Some experimental drugs could be identifiable from their side effects quite independently of its effect on pain (Kirsch 2003). If some of the

participants in the hidden' group had strong suspicions that they were receiving an analgesic, this would tend to enhance the effects of the 'hidden' administration and make it more difficult for the effect of open administration to be greater, and hence to demonstrate a belief effect. If Kirsch's worry is well-founded, then we would expect a reduction in the difference between open and hidden administration. Therefore (again, if Kirsch's worry is well-founded), since the study already provided evidence for a difference between open and hidden administration (and hence expectations effects), we can conclude that the study provides even stronger evidence for expectation effects than is indicated by the results.

6.3.3. Beliefs of the Dispensers: When the 'Pygmalion Effect' is a Confounder

A classic, though non-medical example of how dispenser beliefs may have effects is the 'Pygmalion experiment'[76], carried out by Robert Rosenthal and Lenore Jacobsen. Pygmalion was the name of a Greek artist who sculpted a statue out of ivory and fell in love with it. Subsequently, the statue came to life. Likewise, it is thought that dispensers who seek a particular outcome can influence, perhaps in unconscious or subtle ways, whether it comes about.

In the spring of 1964, in a real public (state funded) elementary school that Rosenthal and Jacobsen call the 'Oak School' (the real name is withheld), experimenters administered the "Harvard Test of Inflected Acquisition" to all (>500) students in grades 1 to 5. Teachers were told that the test "predicts the likelihood that a child will show an inflection point or "spurt" [i.e. point of rapid academic improvement] within the near future" (Rosenthal and Jacobson 1992, vii). Teachers administered this test, but the tests were scored separately by two blind assessors. Then, the teachers were given names of the students who were most likely to "spurt".

> As a reason for their being given the list of names, teachers were told only that they might find it of interest to know which of their children were about to bloom. They were also cautioned not to discuss the test findings with their pupils or the children's parents" (Rosenthal and Jacobson 1992, p.70)

After a year, the same IQ test was administered by the teachers and graded by independent, blind assessors. The "spurters" improved significantly more than the

[76] I am grateful to Dr. Rupert Sheldrake for bringing this example to my attention.

others (see table below). The top 20% of the students named by the test improved in all areas significantly more than the other students (results summarized below)[77].

9. Table 6.1: Mean gain in Total IQ after One Year by Experimental- and Control-Group Children in each of Six Grades[78]

GRADE	CONTROL		EXPERIMENTAL		EXPECTANCY ADVANTAGE	
	N	GAIN	N	GAIN	IQ POINTS	ONE-TAIL $p < .05$*
1	48	+12.0	7	+27.4	+15.4	.002
2	47	+7.0	12	+16.5	+9.5	0.02
3	40	+5.0	14	+5.0	-0.0	
4	49	+2.2	12	+5.6	+3.4	
5	26	+17.5 (+)	9	+17.4 (-)	-0.0	
6	45	+10.7	11	+10.0	-0.7	
TOTAL	255	+8.42	65	+12.22	+3.8	0.02

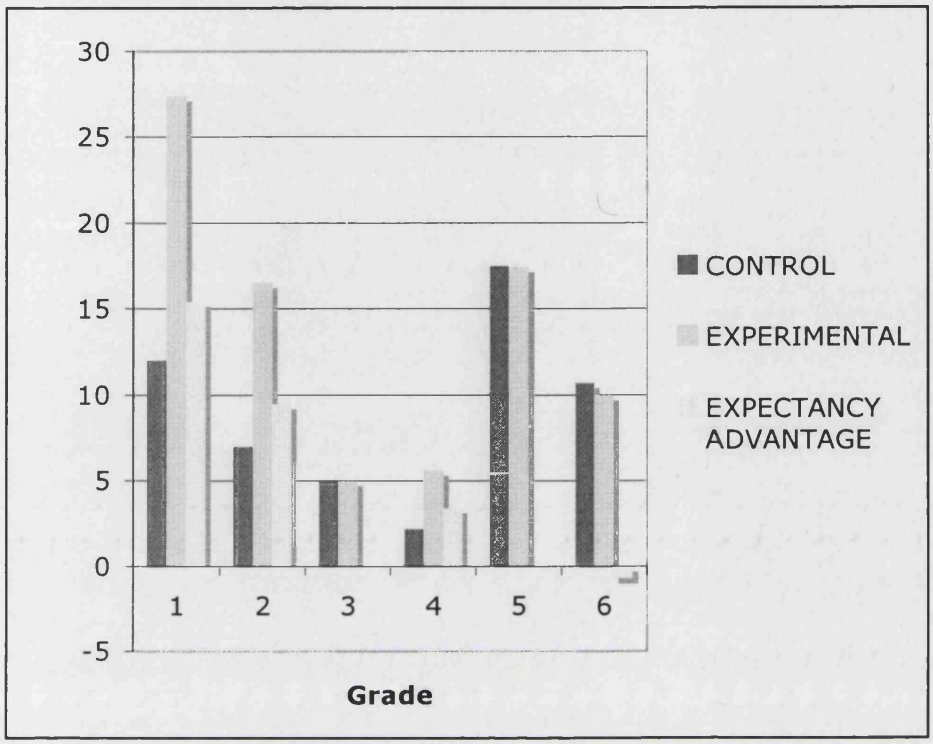
* Mean square within treatments within classrooms = 164.24

The results become more dramatic in a chart (see below).

[77] This is surely a major criticism of the IQ test used since IQ is supposed to be inherent and not subject to environmental influence - the argument would be that even if we assume that this is a case where the expectations of the teachers had a real effect this could not have been on 'real' IQ and if it seemed to be such an effect on this particular test then the test was not a real test of IQ.

[78] Recreated from table 7.1 in (Rosenthal and Jacobson 1992, p. 75). Note that the number of participants in the experimental group is not exactly 20% of the total. This is because "it was felt more plausible if each teacher did not have exactly the same number or percentage of her class listed" (Rosenthal and Jacobson 1992, p.70).

10. Chart 6.1: Effect of Teacher Expectancy Measured as IQ Score Improvement



It is interesting that the "spurters" also scored higher in other tests, such as verbal IQ, reasoning IQ. Even their behaviour and social adaptability was improved relative to the "non-spurters".

In fact the test was a standard IQ test, and the 20% of students who were predicted to "spurt" were chosen completely at random.

The Oak School experiment suggests that the expectations of teachers (and students) can have objective effects on student performance. More generally it suggests that "*one person's expectation for another person's behavior can quite unwittingly become a more accurate prediction simply for its having been made*" (Rosenthal and Jacobson 1992, vii)[79].

---

[79] Interestingly, Popper uses the name *Oedipus Effects* for the same phenomenon as Pygmalion effects (it is unclear whether Popper was familiar with the Pygmalion experiments): "Years ago I introduced the term '*Oedipus effect*' to describe the influence of a theory or expectation or prediction *upon the event which it predicts* or describes: it will be remembered that the causal chain leading to Oedipus' parricide was started by the oracle's prediction of this event. This is a characteristic and recurrent theme of such myths, but one which seems to have failed to attract the interest of the analysts, perhaps not accidentally" (Popper 1969, 1.II, footnote 3).

The mechanism of Pygmalion effects is not necessarily mysterious. A teacher, believing that a student was ready to 'spurt' might pay special attention to that student which could easily translate into accelerated rates of improvement. At the same time, the scarce resources spent on the 'spurters' is not 'wasted' on those less likely to improve.

If there are 'Pygmalion effects' in medicine, then if a dispenser believes that she is administering the best experimental treatment (as opposed to placebo) to a patient, then her belief may translate into improved outcomes of the experimental treatment that have nothing to do with its characteristic[80] features. A caregiver, believing that an extremely ill participant was being given a great new treatment, coupled with the belief that the new treatment is effective, might provide that patient with a higher quality of care. On the other hand, if the caregiver believed that a different patient was being given a placebo, the dispenser might not bother providing the highest quality of care – it might be 'not worthwhile', especially given that they all have scarce resources to distribute amongst their many patients. An obvious scenario where dispenser knowledge could have effects is if the dispenser has an interest in showing that the experimental treatment works. For example, if the intervention administrator discovered the experimental treatment and stands to become famous if it is proven effective, or if she has a financial stake in the experimental treatment then their knowledge of which is the experimental intervention could affect the quality of care they provide. The role of these personal or financial interests could be either conscious or, more charitably, unconscious.

Benedetti's pain study and the Pygmalion study show that at least in some circumstances participant and dispenser beliefs seem to have genuine effects. Double masking would clearly rule out these effects if they are confounding.

The rationale for double masking can therefore be summarized as follows. At least at the start of a double-blind trial, agents are told (and presumably believe) that they have an equal chance of being in the experimental or control group. This prevents the potential effects of participant or dispenser beliefs from confounding the study.

---

[80] The term 'characteristic' is borrowed from Grünbaum (1986) and is used to describe features of a treatment process that are sometimes referred to as 'specific' or 'active'. For example, fluoxetine would be the characteristic feature of treatment involving Prozac. Other features, including the prescription from a sympathetic physician, would be 'incidental'.

I will now examine cases where expectations in the experimental and control group might not be confounding, and therefore where double masking might not add value to a trial.

6.4. Why Double Masking is Less of a Worry for Active Rather than Placebo Controlled Trials

Recall from the second chapter that a confounding factor is one that

4. is unrelated to the experimental intervention,
5. is a determinant of the outcome, and
6. is unequally distributed between experimental and control groups

In this section, I will argue that the worry about patient expectations and dispenser attitudes being unequally distributed in experimental and control groups is not as great in active controlled trials as it is in placebo controlled trials.

If participant beliefs about effectiveness are the same in the two arms of an active controlled trial (and after all neither treatment is presented as a mere placebo) then there would of course be no confounding by them if the trial were open. My expectations regarding aspirin and a new NSAID[81] for mild headache, for example, are likely to be the same and any trial comparing the two would not be confounded by failing to remain single masked.

But are those beliefs likely to be identical in general? Research (Chalmers 1997) suggests that they are not. People seem to believe that *ceteris paribus*, the latest intervention is more effective than the older standard intervention, in spite of the fact that the new interventions are more likely to be *less effective* than the older interventions. In an open active controlled trial, if everyone in the trial believed that the latest intervention was best, then the beliefs and effects of beliefs could well be different in the test and control groups. Or, take an imaginary trial that compared an alternative therapy and a conventional therapy for asthma. If, for some reason, the beliefs about the positive effects of alternative therapies were greater than the beliefs about the positive effects of conventional therapies (perhaps because of the often extended consultations provided by alternative therapies), then knowledge of which group a participant were in could confound the study.

---

[81] Non-Steroidal, Anti-Inflammatory Drug, such as ibuprofen.

I will call beliefs about the effectiveness of treatments that have nothing whatsoever to do with the characteristic effects of the intervention 'prejudice'. Once prejudice is recognized as a potential confounder, it can be controlled for, at least partially. For example, participants with preferences for the experimental intervention could be introduced in equal measure to both groups. In the imaginary trial of a conventional versus alternative treatment, those who preferred alternative therapies could be divided equally among the 'conventional' and 'alternative' groups. Restricted randomization is probably the easiest way to achieve this, but it is also possible to adjust *post hoc* for prejudice even in observational studies.

Or, a more sophisticated way to control for prejudice is to employ the so-called 'patient preference' design. Here

> Patients may be placed in one of three groups according to preference and willingness to be randomised: (*a*) patients who have no strong preferences and therefore consent to randomisation; (*b*) patients with a preference who still consent to randomisation; and (*c*) patients who refuse randomisation and opt for their treatment of choice (Torgerson and Sibbald 1998)

In short, patients with explicit preferences would be given the opportunity to opt out of the randomized trial and receive their preferred intervention.

To be sure, people's 'prejudices' may well not be apparent to them. If not then it will be impossible to control for them. Still, controlling for participants' 'conscious' prejudice will *reduce* any potential confounding effects of prejudice even it is not altogether eliminated.

Moreover it wouldn't be patients' prejudice that would be a concern in an open actively controlled trial of regular (non-alternative) treatment, but rather the beliefs of the investigators (of course then subtly conveyed to the patients). Controlling for patient prejudice at the outset of the trial would not reduce the potential confounding effect of the dispensing investigators. However, even dispenser beliefs could be controlled for, at least partially, in the same way that participant beliefs are controlled for. For instance, dispensers with strong beliefs about the positive effects of the experimental intervention could be balanced out by investigators with strong beliefs about the positive effects of the standard control. Then, of course, masked *assessment* should be employed wherever possible especially if the assessment is performed by the dispensers.

These strategies for reducing confounding of participant expectation and dispenser attitudes, although not perfect, are nonetheless surely useful. These strategies

will not generally be successful for placebo controlled trials, where it would be difficult to find any participants at all who have a prejudice for placebo treatment.
Prejudice may be exacerbated or caused by what I will call hype. Using Jefferson's word quoted at the outset of this chapter, some treatments are more fashionable than others. This could be due to the fact that the older, standard treatments have accumulated an extensive side effect profile while the new treatment represents hope. This hope is sometimes boosted by aggressive marketing. For example, *Prozac*, an SSRI antidepressant, was marketed to consumers in the United States before it was approved for marketing (a practice illegal outside the United States and New Zealand). Eli Lilli, the company who developed Fluoxetine, chose the name *Prozac*, which was unrelated to the characteristic chemical compound that was supposed to be responsible for the effect on serotonin levels. Previous to *Prozac*, drug names were often related to the characteristic chemical. The name *Prozac* was probably chosen because of the positive connotations of its prefix. The end result was that the number of 'depressed' people demanding prescriptions (some even before they were approved for marketing) for SSRIs such as *Prozac* skyrocketed to levels unheard of in the past (Block 2007; Healy 2004, introduction).

Two other historical facts are worth noting. First, the side effects of the older class of antidepressants, known as tricyclics, that had been developed in the late 1950's, were well known. Second, depression as a disease was becoming more widely recognized, some argue because of the marketing of depression by the same companies that manufactured the new generation of antidepressant drugs (Gardner 2003). As a result, more people were thinking of themselves as depressed, and they had good reason to be wary of the older drugs. It would be reasonable to expect an open active controlled trial of *Prozac* versus an older tricyclic to be biased because people had been persuaded to believe that *Prozac* was (potentially) a much better option than the older treatment. If the vast majority of potential participants in a trial had been 'hooked', then it would be difficult to have an equal number of Prozac and standard treatment supporters in each group. However assuming that there were at least some who were not 'hooked' on Prozac before the trial began, it would be possible to adjust, *post hoc*, for this potential confounder.

Note that although I have discussed the case where the new, 'fashionable' treatment has been hyped and that this hype may have effects that have nothing whatsoever to do with the characteristic features of the treatment, the same phenomenon

could occur with an older intervention. For simplicity, I will continue to discuss the more common case where the hype has the potential to exaggerate the effects of the new experimental intervention.

Further, although it is difficult to control for the inevitable effects of hope that some new intervention will be more effective or less harmful than an older intervention or than no intervention without double masking or placebo controls, it is important to separate this hope from unjustified advertising of new interventions. If the advertising is conducted before there is sound evidence to support the view that that new interventions are safe and effective[82], then it is false that we have any grounds to believe that an experimental treatment is more effective than an older one.

Indeed there are equally good reasons to fear new treatments as there are to expect that they will be better. Because early trials are usually relatively short, or at any rate often too short to pick up long-term side effects, there are as many reasons to *fear* newer experimental treatments as there are to hope that they are better than standard care.

The problem with hype that I have described is thus partly a problem with direct to consumer advertising (DTA) as currently practised, either before or after approval for marketing. Surely the just thing to do is to provide complete information about all available alternatives to consumers. Nonetheless, in cases where hype surrounds a new treatment, it is surely best to attempt to keep the trial double masked or to carefully adjust for this potential confounder.

To sum up this section, there are good reasons to believe that participant expectation and dispenser attitudes have less potential for confounding an active controlled trials than they do in placebo controlled trials. This is because the choice in an active controlled trial is between two potentially effective non-placebos, rather than between potentially effective non-placebo and placebo. Nonetheless, people often seem to have greater expectations regarding the effectiveness of the newer treatment, in spite of the fact that newer treatments are not usually more effective. In some cases these

---

[82] I leave a discussion of whether this trust is justified to another study. Suffice it to note that at least in the case of the FDA, the bias is towards approval of the new treatments. Indeed until 1962 a new drug merely had to demonstrate *safety*, and not effectiveness. They still do not need to demonstrate greater effectiveness than the best existing treatment.

differential beliefs can be controlled for explicitly at the outset of the trial, or in *post hoc* adjustment.

Another circumstance where participant expectations and dispenser beliefs may not confound a study as much as we think is where these potential confounders have no actual effects. I will consider this case now.

6.5. Where Participant Expectation and Dispenser Attitude Do Not Confound a Study

In this section I will outline cases where participant and dispenser beliefs do not affect the outcome of the study in a way that can be called confounding. These situations are where the belief effects are on the 'causal pathway' of the characteristic features, and any trial where the characteristic effects swamp the possible effects of participant and dispenser beliefs.

Even if the effects of beliefs are different in the test and control groups, if beliefs are characteristic rather than incidental features[83], then they cannot be considered confounders. To see why, imagine there were a drug whose only characteristic feature was a chemical that was dramatically effective at making any depressed person who took it *believe* that the intervention had powerful characteristic effects that cured depression. Call the characteristic feature of this new drug the *x*-factor. Because depression is probably particularly sensitive to beliefs, the drug containing the *x*-factor may well prove very effective for the treatment of depression. However the drug has no *other* characteristic features for changing the chemicals, such as serotonin, that are currently believed to be correlated with, or cause, depression. Imagine that this drug demonstrated significantly superior effectiveness to standard SSRI antidepressants. The only reason that the new drug was more effective was because of participant belief. Adopting the rule that *all* belief effects are confounding would lead one to claim that the

---

[83] I am referring to the characteristic/incidental distinction originating from Grunbaum (1986) and discussed in chapter 3. According to my earlier discussion this distinction is made by an underlying therapeutic theory. However for present purposes what I mean by 'characteristic' and 'incidental' matches the more common 'active' or 'specific' used in the medical literature to distinguish between placebogenic and non-placebogenic features of a treatment process. For instance, fluoxetine in a Prozac pill will be the only characteristic feature in treatment with a Prozac pill. The incidental features in this case might be the manner in which the pill is delivered, etc.

imaginary study was confounded by the different beliefs even though the beliefs were a direct result of the characteristic feature. In short, if the increased expectations in the experimental group arise from the characteristic features of the experimental treatment, then they cannot be considered confounding.

Participant and dispenser beliefs might not be a worry where their potential confounding effect is *large* relative to the size of the characteristic effects of the test intervention. Acute appendicitis or meningitis might well be influenced by beliefs, but it is unlikely that the effects of belief are significant relative to the effect of the treatment. For this reason, appendectomies and antibiotics for meningitis have of course never been tested in double masked conditions. The effects of a participant believing he is receiving an effective intervention may well have some effects, but it is unlikely that these effects would be strong enough to explain avoiding death from acute appendicitis. As Smith and Pell imply, you don't need a double masked RCT to know that parachute use prevents death in someone falling from a plane (Smith and Pell 2003).

The case where the characteristic effects appear dramatic explains the Phillip's Paradox stated at the outset of this chapter where dramatically effective treatments cannot be tested in double blind conditions. At least according to the view that double masked studies are of higher quality than open studies, paradoxically, dramatically effective treatments, are not supported by the best possible evidence. However where the treatment effect is dramatic (such as appendectomy for acute appendicitis), it is safe to assume that expectations or attitudes could not account for the entire effect. Hence, although participant expectations and dispenser attitudes might have confounded the study, they were not sufficiently powerful to present a rival hypothesis for the entire effect. In short, the potentially confounding effects of expectations and attitudes in trials of dramatically effective treatments are relatively insignificant. Therefore, the Phillip's Paradox dissolves once we abandon universal adherence to the rule that double masking increases quality and instead evaluate studies based on 'scientific common sense', namely the view that good evidence rules out plausible rival hypotheses.

In yet another circumstance, participant and dispenser beliefs might not have significant effects; if not, then double masking will not add methodological value. The interesting studies of participant expectation or belief conducted by Benedetti's team, along with the Pygmalion studies and studies of different coloured placebos, might lead one to believe that participant and dispenser attitudes can, and often do, play a significant role. However, the dispenser, or teacher effects in the Pygmalion studies

tapered off as the students aged (see table 6.1). Likewise, Benedetti's study, although it showed significant effects of participant belief, was restricted to studies of a few ailments and did not show clinically relevant effects. Current evidence suggests the magnitude of participant and dispenser beliefs varies quite widely. In some cases, they might not have any effects at all while in others they may well have clinically relevant effects. Double masking will help in the latter, but not the former cases. The structure of my argument is as follows:

(1) If there are participant and dispenser beliefs then there are placebo effects[84].

(2) There is evidence that placebo effects are insignificant outside treatments whose outcomes are subjective and continuous.

(3) Therefore, there is evidence that participant and dispenser belief effects are insignificant in some cases (such as for certain objective outcomes).

If we identify the set of incidental features what produces any overall placebo effect that there may be in particular circumstance, then if participant and dispenser beliefs have effects it follows that there are *placebo* effects – since no 'regular' therapeutic theory gives a 'characteristic' role to such beliefs (at any rate in regular pharmacological treatments, where the characteristic features are exclusively the 'active' chemicals). In two recent meta-analyses, Asbjørn Hróbjartsson and Peter Gøtzsche (2001) looked at 3-armed trials that included experimental, placebo control, and untreated groups and found no overall significant placebo effect. If we assume that the three groups were always free from selection bias, that the "untreated" groups were actually untreated, and the placebo controls were legitimate, the placebo effect could fairly be estimated as the difference between the average effect in the placebo group less the average effect in the untreated group.

Defining placebos "practically as an intervention labelled as such in the report of a clinical trial" (Hróbjartsson and Gøtzsche 2001, p. 1595) Hróbjartsson and Gøtzsche searched several major medical databases[85] for 3-armed RCTs. They excluded studies where participants were paid or were healthy volunteers, where the outcome assessors

---

[84] In theory, if participant and dispenser belief effects are offset by the effects of other incidental features, there would be no overall placebo effects even if there were participant and dispenser belief effects. However in practice this is unlikely to be the case.

[85] Medline, EMBASE, PsychLIT, Biological Abstracts, and the Cochrane Controlled Trials Register up to 1998.

were unmasked, where the dropout rate exceeded 50%, and when "it was very likely that the alleged placebo had a clinical benefit not associated with the ritual alone (e.g. movement techniques for postoperative pain)" (Hróbjartsson and Gøtzsche 2001, p. 1595). (I will discuss whether their exclusion criteria were justified after outlining the study.)

After identifying 727 potentially eligible trials, they excluded 404 for not being randomized, 129 for failing to have a placebo group or an untreated group (although they were described as having one), 29 for being reported in more than one publication, 11 for using unmasked outcome assessment, 24 for meeting other exclusion criteria such as high dropout rates. Sixteen trials did not include relevant outcome data. This left 114 trials for the meta-analysis. Typical pill placebos were lactose pills, typical 'physical' placebos were procedures performed with the machine turned off (e.g. sham transcutaneous electrical nerve stimulation), and typical psychological placebo was theoretically neutral discussion between participant and dispenser. Over 40 clinical conditions were included in the analysis, ranging from hypertension and compulsive nail biting to fecal soiling and marital discord.

They classified the trials according to whether the outcomes were binary ("yes" or "no") or continuous, and whether the outcomes were subjective or objective. For binary outcomes, they calculated the relative risk of an unwanted outcome, which is the ratio of the number of participants with an unwanted outcome to the total number of patients in the placebo group divided by the same ratio in the untreated group. A relative risk below 1 therefore indicates a positive placebo effect. With continuous outcomes, the authors calculated the standard mean difference, the difference between the mean value for an unwanted outcome in the placebo group and for the no treatment group, divided by the pooled standard deviation. A value of −1 indicates that the mean in the placebo group was 1 standard deviation below the mean in the untreated group. The results are summarized in the two tables below.

11. Table 6.2: Effect of Placebo in Trials with Binary or Continuous Outcomes (From Hrobjarsson and Gøtzsche 2001, p. 1596)

Outcome	No. of Participants	No. of Trials	Pooled Relative Risk (95% CI)†
Binary			
Overall	3795	32	0.95 (0.88 to 1.02)
Subjective	1928	23	0.95 (0.86 to 1.05)
Objective	1867	9	0.91 (0.80 to 1.04)
			Pooled Standardized Mean Difference (95% CI)‡
Continuous			
Overall	4730	82	−0.28 (−0.38 to −0.19)
Subjective	3081	53	−0.36 (−0.47 to −0.25)
Objective	1649	29	−0.12 (−0.27 to 0.03)

*CI denotes confidence interval.

†The relative risk was defined as the ratio of the number of patients with an unwanted outcome to the total number of patients in the placebo group, divided by the same ratio in the untreated group. A value below 1.0 indicates a beneficial effect of placebo.

‡The standardized mean difference was defined as the difference between the mean values for unwanted outcomes in the placebo and untreated groups divided by the pooled standard deviation. A negative value indicates a beneficial effect of placebo.

Although there was significant (p=0.003) heterogeneity among trials with binary outcomes, placebo did not have a significant effect... (overall pooled risk of an unwanted outcome with placebo 0.95; 95 percent confidence interval, 0.88 to 1.02). For continuous outcomes, there was a significant placebo effect for trials with subjective outcomes (-0.36; 95% confidence interval –0.47 to - 0.25), but not for trials with objective outcomes (-0.12; 95% confidence interval –0.27 to 0.03). There was also significant heterogeneity (p=0.001) for trials with continuous outcomes. However, there was significant (p=0.05) relationship between size of trial and placebo effect, indicating that bias due to small trials played some role. Of all the ailments that were treated in more than 3 trials, only pain showed a significant effect (-0.27; 95% confidence interval –0.40 – 0.15 - see table below).

12. Table 6.3: Effect of Placebo on Specific Clinical Problems (From Hróbjartsson and Gøtzsche, 2001, p. 1597)

Outcome	No. of Participants	No. of Trials	Pooled Relative Risk (95% CI)†
Binary			
Nausea	182	3	0.94 (0.77 to 1.16)
Smoking	887	6	0.88 (0.71 to 1.09)
Depression	152	3	1.03 (0.78 to 1.34)
			Pooled Standardized Mean Difference (95% CI)‡
Continuous			
Pain	1602	27	−0.27 (−0.40 to −0.15)
Obesity	128	5	−0.40 (−0.92 to 0.12)
Asthma	81	3	−0.34 (−0.83 to 0.14)
Hypertension	129	7	−0.32 (−0.78 to 0.13)
Insomnia	100	5	−0.26 (−0.66 to 0.13)
Anxiety	257	6	−0.06 (−0.31 to 0.18)

*Only problems addressed by at least three trials are included. CI denotes confidence interval.

†The relative risk was defined as the ratio of the number of patients with an unwanted outcome to the total number of patients in the placebo group, divided by the same ratio in the untreated group. A value below 1.0 indicates a beneficial effect of placebo.

‡The standardized mean difference was defined as the difference between the mean values for unwanted outcomes in the placebo and untreated groups divided by the pooled standard deviation. A negative value indicates a beneficial effect of placebo.

In sum, there were significant placebo effects for trials of pain (where the outcome measure was subjective) and in general for trials of ailments with subjective continuous outcomes. Even in these cases Hróbjartsson and Gøtzsche question whether these studies provide evidence of placebo effects.

> Patients in an untreated group would know they were not being treated, and patients in a placebo group would think they were being treated. It is difficult to distinguish between reporting bias and a true effect of placebo on subjective outcomes, since a patient may tend to try to please the investigator and report improvement when none has occurred. The fact that placebos had no significant effects on objective continuous outcomes suggests that reporting bias may have been a factor in the trials with subjective outcomes" (Hróbjartsson and Gøtzsche 2001, p. 1597).

The so-called 'Hawthorne Effect', which is part of what we mean when we talk about placebo effects, is, very briefly, the positive effect of being in an experiment no matter

what the intervention is.[86] If there are Hawethorne Effects, then we might expect them to be greater in the placebo control group than in the 'no treatment' group. If so, then we would expect the placebo effect to be enhanced relative to no treatment.

Hróbjartsson and Gøtzsche conclude that there is:

> little evidence that placebos in general have powerful clinical effects. Placebos had no significant pooled effect on subjective or objective binary or continuous objective outcomes. We found significant effects of placebo on continuous subjective outcomes and for the treatment of pain but also bias related to larger effects in small trials. The use of placebo outside the aegis of a controlled, properly designed clinical trial cannot be recommended (Hróbjartsson and Gøtzsche 2001, p. 1599).

There are several problems with Hróbjartsson and Gøtzsche's meta-analysis. First, it can be questioned whether the statistical results of the study warrant their concluding statements. Overall, there was a significant placebo effect in trials with continuous outcomes, and there were more participants in trials with continuous outcomes than in trials with binary outcomes. This would surely have tempted some authors to claim that there are often significant placebo effects, rather than conclude that there is little evidence for significant placebo effects.[87]

Then, Hróbjartsson and Gøtzsche's warning that significant placebo effects for pain and subjective continuous outcomes may be due to bias is contrived. If treated in a way that made them feel they were in a trial, the 'untreated' group may have experienced Hawthorne Effects as well, in which case the apparent placebo effects would be reduced. Hróbjartsson and Gøtzsche admit that if participants in the 'untreated' groups sought treatment outside the study, the apparent placebo effects would have been reduced (Hróbjartsson and Gøtzsche 2001, p. 1597). In short,

---

[86] The Hawthorne Effect is named not after an author, but after the name of a series of experiments in the Hawthorne works of the Western Electric Company in Chicago between 1924 and 1933. In one study, for example, the illumination remained stable for the control group and was slowly raised for the experimental group. Productivity increased equally in both groups. In another study, illumination was stable in the control group but was slowly lowered in the experimental group. Productivity increased steadily and equally in both groups until the lights were so low in the experimental groups that the experimental workers protested and production fell off (Roethlisberger and Dickson 1939). The "Hawthorne Effect" has been interpreted many ways. Here I will take it to be the potential (sometimes temporary) positive effects of being in an experiment.

[87] Masking the assessors and manuscript writers may have prevented this possible bias.

Hróbjartsson and Gøtzsche's claim that bias may have affected the studies could work either to enhance or reduce the estimate of the true placebo effect, and not simply to reduce it as they suppose.

Further evidence of Hróbjartsson and Gøtzsche's bias is evident elsewhere. They state that: "It surprised us that we found no association between measures of the quality of a trial and [increased] placebo effects" (Hróbjartsson and Gøtzsche 2001, p. 1598). One could just as easily be surprised to find no association between measures of the quality of a trial and *decreased* placebo effects. Poor trial quality is usually associated with exaggerated treatment effects. One way to exaggerate treatment effects is for the placebo effect to decrease. On the other hand, increase of trial quality could mean that a trial is visibly (to the participant) more tightly controlled. This increase in control could lead to increased Hawthorne Effects, and hence increased placebo effects. In short, the quantitative results of the meta-analysis do not support the conclusion that there are no placebo effects.

Third, Hróbjartsson and Gøtzsche imposed insufficient normative constraints on what they took to be placebos and 'no treatment', which may have led to erroneous results. Recall that the estimate of the placebo effect is based on the average difference between 'placebo' and 'no treatment'. But if the placebo controls were illegitimate, i.e. they had characteristic features or they didn't have all the potentially effective incidental features, then the estimated 'placebo' effect could be erroneous. Or, if what they counted as 'no treatment' in fact amounted to some form of placebo or treatment, then the 'placebo effect' will have been underestimated. A superficial glance at what the authors counted as 'placebos' or 'no treatment' controls suggests that their criteria for what counted as 'placebo' controls or 'no treatment' may well have led to mistaken results.

Initially the authors attempt to skirt around the thorny issue of defining placebos and claim to define placebos as whatever is called a placebo in a report of a clinical trial. Yet later they exclude the trial if "it was very likely that the alleged placebo had a clinical benefit not associated with the ritual alone (e.g. movement techniques for postoperative pain)" (Hróbjartsson and Gøtzsche 2001, p. 1595). That is, they rightly excluded the trial if the placebo control was illegitimate because it included characteristic features of the test treatment (see chapter 5). Although they can be forgiven for giving up their initial claim that they would define placebos 'practically' as whatever was used as a placebo control in a trial, the fact that they had no coherent rules

for deciding what counts as a legitimate placebo made their choice unsystematic and possibly biased. In fact, they jumble (along with placebo pills and injections) relaxation (described as a placebo in some studies and a treatment in others), leisure reading, answering questions about hobbies, newspapers, magazines, favourite foods and sports teams, talking about daily events, family activities, football, vacation activities, pets, hobbies, books, movies, and television shows as placebos (Kirsch 2002). In short, although their exclusion criteria are legitimate, they need to go further before we can accept that the placebos they examined are legitimate. If they are not, then the results of the meta-analysis cannot be relied upon.

A parallel problem that has yet to be pointed out in the literature is the failure to impose restrictions on the untreated groups. If the 'untreated' groups did something having a clinical benefit, this would reduce the difference between 'no treatment' and placebo and hence the estimated placebo effect. On the other hand, if the 'untreated' participants were closely monitored, then placebo effects, such as Hawthorne Effects could have resulted. Either way, the effects of being left 'untreated' may have been exaggerated, which would have led to an *under*estimation of placebo effects.

Further, the highly significant heterogeneity of the interventions studied calls into question what conclusions can be drawn from the meta-analysis. An overall finding of insignificant placebo effects does not count against the reasonable view that placebo effects are common and powerful for certain disorders (those whose outcomes are measured subjectively), but not for others. Indeed had the Hróbjartsson and Gøtzsche study used masking of the data collectors, statisticians, and manuscript writers it is unclear whether their conclusion, or the title of their work would have been the same.

There is also a suggestion implicit in their work that 'subjective' means 'unreal'. An outcome such as decreased pain can be very important clinically even though it can only be measured (at any rate directly) 'subjectively'. This also ties in with what they call bias. Although 'bias' in the reporting of subjective outcomes may be a real worry, it is not even clear to what extent it is possible to make sense of claims that a patient reports decreased pain to 'please' the investigator, but allegedly, in fact, is not experiencing 'real pain'. In short, the subjective nature of the outcome does not make it 'unreal'.

In reaction to criticisms, Hróbjartsson and Gøtzsche dug in their heels. They updated the 2001 meta-analysis in 2004, in a study which supposedly confirmed the results of the earlier meta-analysis (Hróbjartsson and Gøtzsche 2004a). Then, in 2006

they reviewed several meta-analyses of apparent placebo analgesic effects and concluded that the analyses had been poorly done. In the meta-analysis update, Hróbjartsson and Gøtzsche recognize the problem that the 'untreated' participants may have been treated: "Patients in a no-treatment group also interact with treatment providers, and the patients are therefore only truly untreated with respect to receiving a placebo intervention" (Hróbjartsson and Gøtzsche 2004a, p. 97). They also mention the problem of heterogeneity: "we cannot exclude the possibility that in the process of pooling heterogeneous trials the existence of such a group [of trials that showed a significant placebo effect] was obscured" (Hróbjartsson and Gøtzsche 2004a, p. 97).

In spite of recognizing at least a few of the problems with their analysis, their conclusions about apparent placebo effects were not at all tempered. The conclusion of the updated paper was similar to that of the first:

> In conclusion, we reproduced the findings of our previous review and found no evidence that placebo interventions in general have large clinical effects, and no reliable evidence that they have clinically useful effects. A possible effect on patient-reported continuous outcomes, especially on pain, could not be clearly distinguished from bias (Hróbjartsson and Gøtzsche 2004a, p. 98).

Problems with the Hróbjartsson and Gøtzsche studies notwithstanding, there are two important points they highlight. First, their method, and not Beechers, for measuring the magnitude of belief effects is correct, if done properly. Measuring the *difference* between placebo and no-treatment, is the correct way to measure placebo effects. Second, it is possible, indeed probable, that the placebo effect varies depending on the intervention, the ailment, and the type of outcome.

Another empirical study needs mentioning before concluding this section. In a study of 33 meta-analyses Schulz *et al.* claim that the unmasked studies odds ratios were exaggerated by 17%.[88] This study seems to suggest that participant expectation

---

[88] Odds ratio = odds of outcome event / odds of no even in control group
odds of outcome event / odds of no outcome event in experimental group

	Outcome event		Total
	Yes	No	
Control Group	a	b	a + b
Experimental Group	c	d	c + d

= (a/b) / (c/d)

= ad / bc

and dispenser attitudes (whose effects were supposedly neutralized in the masked studies) have a 17% effect. I will leave aside the fact that other studies of the effects of unmasking have had conflicting results (Moher et al. 1998; Miller, Colditz, and Mosteller 1989) to focus on a problem they all share. Schulz's team failed to define which groups were masked. If, for example, it was the outcome assessors and not the dispensers that were masked in the masked studies and unmasked in the unmasked studies, their beliefs could have been responsible for the different effects. The authors recognized this shortcoming, admitting that the meager information on the approaches used for double masking made it difficult to interpret what the source of exaggerated effects are (Schulz et al. 1995). In light of the ambiguity with which the term double masking has been used, it is difficult to interpret this study as evidence for participant and dispenser belief effects.

To sum up this section, double masking will not reduce the potentially confounding effects of participant and dispenser beliefs as much in active controlled trials as it will in placebo controlled trials. Then, there are cases where beliefs are not confounders. These cases are (a) where the beliefs are characteristic features of the treatment, (b) where the belief effects are small relative to the characteristic effects of the experimental treatment, and (c) where there are no belief effects.

I will now question whether double masking is a realistic goal.

6.6. The Near-Impossibility of *Successful* Double Masking

Attempting to double mask a study is one thing; keeping it successfully double masked for the duration of a trial is another. There is strong evidence to suggest that even when the best efforts are made, double masking is rarely successful for the duration of the trial. This is because to keep the trial successfully masked means that the appearance, smell, taste, and side-effects of the experimental intervention must be mimicked by the control. Otherwise, given the (usually enforced) requirement of full informed consent whereby participants are made aware of the nature of the experimental intervention (potential effects, side effects, etc.), participants and dispensing physicians will correctly guess whether a particular intervention is the experimental treatment or

---

The odds ratio is a measure of effect size. An increase in 17% of the odds ratio, however, does not mean an increase of 17% absolute effect size. A 17% increase in odds ratios can be due to a much smaller increase in absolute effect size.

control. If it is the case that double masking is unlikely to be successful, then the apparent advantage of being double masked' is nullified and open trials are equally good.

In a recent study, Fergusson, Glass, Waring, and Shapiro (2004) investigated whether trials described as double masked were successful. Whether a trial is successfully double masked can reasonably be ascertained by asking participants and dispensers to guess which participants were in the treatment or control groups –if guesses do significantly better than chance, then there is a degree of 'unmasking'. These authors conducted a Medline search of randomized, placebo controlled trials published from 1998 to 2001 in 5 top general medical and 4 top psychiatry journals[89]. Their search turned up a total of 473 medical trials and 192 psychiatry trials. From this group they randomly selected 100 trials in each group. Nine of the randomly selected trials were excluded because they were not placebo controlled in spite of being described as such[90]. They ended up with 97 medical trials and 94 psychiatry trials.

Of the 97 medical trials, only 7 provided evidence of the success of double masking. Of those, 5 reported that the masking was unsuccessful[91]. Of the other two, one described study claimed blinding as successful without further comment or statistical data (Fergusson et al. 2004). In short, even where masking is tested-for, there was insufficient information provided to verify the results.

Of the 94 psychiatry trials, 8 reported evidence of testing for successful masking. Four of these reported that the masking was unsuccessful. Overall:

> 15 of the 191 trials (8%) provided such information, be it qualitative or quantitative. Of the 15 trials, only five trials reported that blinding was

---

[89] These were the *Journal of the American Medical Association (JAMA), New England Journal of Medicine (NEJM), the British Medical Journal (BMJ), Annals of Internal Medicine, Archives of General Psychiatry, Journal of Clinical Psychiatry, British Journal of Psychiatry,* and the *American Journal of Psychiatry.*

[90] In spite of noting that double masking is important for active controlled trials, the authors limited their evaluation to placebo controlled trials.

[91] The authors don't define success. I will assume that it means that significantly more participants correctly guessed which group they were in than would be predicted by chance alone.

> successful and of these, three did not present any quantitative data analysis to support their claim (Fergusson et al. 2004)[92].

This study suggests that masking is rarely tested for and, where tested for, rarely successful. It incited a flurry of responses on the BMJ website by prominent medical researchers such as Stephen Senn, David Sackett and Douglas Altman. Describing the well-known 'Tea Lady Experiment' Senn claim that 'unsuccesssful' masking is not a methodological failing as long as it is a result of the efficacy of the treatment:

> The classic description of a blinded experiment is Fisher's account of a woman tasting tea to distinguish which cups have had milk in first and which cups have had tea in first in support of her claim that the taste will be different. Here the efficacy of the treatment, order of milk or tea, is "taste" and the lady's task is to distinguish efficacy. Fisher describes the steps, in particular randomization, that must be taken to make sure that the woman is blind to the treatment. But if he were to adopt the point of view of Fergusson et al, there would be no point in running the trial, since if the lady distinguished the cups, the trial would have been described as inadequate, as she has clearly not been blind throughout the trial. (Senn 2004)

What Senn claims is that, in cases where the participant (i.e. the 'Tea Lady') correctly guess which is the 'experimental intervention' (i.e. the milk first cups of tea) *because they have identified the characteristic feature* (i.e. whether milk was poured first), that it is mistaken to call the trial inadequate.

Although Senn may be mistaken about the classic description of a blinded experiment – the point of the Tea Lady experiment is to demonstrate the value of concealed random allocation – his point that successful masking is difficult to achieve with effective treatments is well made. Sackett echoes Senn's view that testing for successful masking once a trial is underway is difficult:

> Once a trial is under way, I'm not smart enough to separate the effectiveness of blinding from the effects of pre-trial hunches about efficacy, and I've never met anyone who is. Accordingly, we rigorously test for blindness before our trials, but not during them and never at their conclusion (Sackett 2004).

Sackett and Senn are surely correct that in some cases, namely where the characteristic features of the test treatment are so apparent that the trial becomes unmasked, should not count against the methodological value of the trial. On this view, investigators should attempt to make a double masked trial at the outset. Then, if the participants or dispensers correctly guess which intervention they are taking or giving, it

---

[92] Fergusson *et al.* proceed to question whether the studies which provided quantitative data on the success of double masking were methodologically sound, and claim that they weren't.

is unimportant because it is often the characteristic features of the treatment that cause the unmasking.

Yet even if Sackett and Senn are correct, their arguments only apply in cases where the unmasking is due to the characteristic features of the experimental treatment. As I explained earlier, if the effects of the experimental treatment are dramatic, then the 'unmasking' of the study cannot be said to have impaired its methodological quality. Yet, unmasking can occur because of many reasons other than characteristic effectiveness, dramatic or otherwise. For instance, cases where the side-effects of the experimental treatment are identifiable will lead to unmasking that does impair a study's methodological quality. Where unmasking occurs for these other reasons, it does in fact call the adequacy of the trial into question. Further, both Sackett and Senn are incorrect that the different reasons for unmasking cannot be disentangled. Perhaps even more relevantly, ignoring other possible reasons for unmasking is unjustified. As Shapiro notes, "[i]t seems contrary to an evidence based approach to avoid obtaining data because we have to struggle with its interpretation. (Shapiro 2004).

To illustrate, imagine a trial of an antidepressant drug whose characteristic chemical has no effects yet that has unique and recognizable side effects versus inactive placebo (one that does not mimic the side-effects of the experimental treatment). If this is the case then participants could guess when they are in the experimental arm of such a trial, which could lead to increased beliefs and expectations about recovery and hence better effects.

In order to determine whether unmasking was due to the characteristic or non-characteristic features, the tests for successful masking could ask *why* participants believe that they were in a particular group. In short, if a trial comes unmasked for reasons other than characteristic effectiveness then it is legitimate to claim that double masking has not achieved the purpose of ruling out confounding participant and dispenser beliefs. Contrary to what Sackett implies, it *is* possible, at least to some extent, to discover the reasons for unmasking.

Another study was conducted earlier this year by a research team at the Nordic Cochrane Centre that made the case for the difficulty of keeping a trial successfully double masked even stronger. Hróbjartsson and colleagues randomly selected a sample of 1599 clinical trials from the Cochrane Central Register of Controlled Trials that were published in 2001. Thirty-one (2%) reported tests of the success (or otherwise) of masking. It was considered successful in 14 of the 31 studies, unclear in 10, and

reported as unsuccessful in 7. The possibility of a biased result due to the unmasking was not addressed or dismissed in 6 of the 7 unsuccessful cases. In short, 2% of the trials were checked for double masking, and of the 2%, less than half reported double masking success.

To test whether the apparent failure to successfully double mask was a problem with reporting, Hróbjartsson et al. selected a random sample of 200 trials not reported as having conducted tests for the success of masking, and asked the authors whether unreported tests had been done. Of the sample, 130 (65%) responded, and 15 (less than 12%) of these reported having conducted, but not reported, the tests. In short, masking was *reported* as successful in a very small proportion – perhaps as low as 1% - of 1599 trials. The authors conclude that "Blinding is rarely tested. Test methods vary, and the reporting of tests, and test results, is incomplete. There is a considerable methodological uncertainty about how best to assess blinding, and an urgent need for improved methodology and improved reporting" (Hróbjartsson et al. 2007). Trials that perform tests of the success of blinding but do not report them must be viewed with suspicion: successful double masking is something to be proud of. This study, combined with the earlier one suggest either that tests for successful double masking are rarely made, and that when made are often methodologically problematic. Most importantly, when tested for and reported, the results in a majority of cases reveal that double masking has been unsuccessful.

To make matters worse, there are good reasons to believe that the problem of successfully double masking a study runs deep. The question of whether studies are successfully double masked has been lurking in the background since I laid down the conditions required for legitimate placebo controls (see last chapter). Recall that the placebo control, in order to gain legitimacy, had to be similar in superficial appearance (taste, smell, touch), side effects, and effectiveness to the experimental treatment. The practical difficulty in designing placebo controls that meet these conditions, especially with regards to obscure side effects, may be insurmountable.

For example, one side effect of SSRIs is increased probability of sexual dysfunction. How could this side effect be imitated? Even if there were some drug which raised the probability of sexual dysfunction that could be added to the control treatment, a further set of problems emerges. First, it would have to be established that the drug did not have effects (either positive or negative) for the target disorder. Second, it would be ethically problematic to deliberately cause harm with a control treatment.

Third, there would still be the other side effects of the experimental intervention (most drugs have more than one side effect) *and* of the drug for sexual dysfunction.

Irving Kirsch and Guy Sapirstein (Kirsch and Sapirstein 1998) investigated whether the characteristic features of *Prozac* and other SSRIs were due to the *belief* that SSRIs are effective, or to the characteristic (pharmacological) features of SSRIs themselves. They first noted that the side effects of SSRIs are different from the side effects of other drugs and placebos. Because of informed consent, participants in a trial are aware of these potential side-effects associated with the experimental and control treatments. Therefore, in spite of attempts to keep the trials double masked, participants often correctly suspect (and are in effect encouraged to suspect) when they are taking the experimental treatment.

In one example, Irving Kirsch and Guy Sapirstein compared the effect of established antidepressant drugs versus inactive placebo with other drugs versus inactive placebo for depression (Kirsch and Sapirstein 1998, p.7). They found that the other drugs (amylobarbitone, lithium, liothyronine, and adinazolam), which had no known characteristic antidepressant features, were at least as effective as established antidepressant drugs. This study could be interpreted to mean that, if we had a 'complete' therapeutic theory, we would be able to identify characteristic features of the 'other drugs' that have positive effects on depression. Or, the hypothesis preferred by the authors, that the non-antidepressant physiological properties (side-effects) of these drugs lead to unmasking of the study and that the unmasking of the study was what accounted for the superior effectiveness of the test treatment over placebo. The participants, suffering from side effects of the experimental treatment (which they knew about because of fully informed consent or their own research), correctly guessed that they were receiving the experimental intervention as opposed to placebo, which led to increased expectations regarding recovery. Correspondingly, participants not suffering from side effects correctly deduced that they were taking the 'placebo' and had lower expectations regarding recovery. Hence, the increased benefit of the experimental treatment could have been due to these differing expectations: "these medications function as active placebos" (Kirsch and Sapirstein 1998, p.7). Although supported by some other independent studies (Moncrieff 2003), the study has not gone uncriticised (Klein 1998). I will not go into a detailed critique of the Kirsch/Sapirstein study here. The idea to be gleaned is that it is important for placebo control treatments to mimic the

side-effects of the treatments they are supposed to imitate. Otherwise, even if a study is initially carefully blinded, it may later, or more often sooner, come unmasked.

It could still be maintained that because it is possible in some cases, trials that attempt to keep the double mask are superior to those that do not. Further, because the problem with *retaining* the double mask has only been highlighted recently, we might expect the rate of successful double masking to increase in the near future. Since double masking raises the probability of ruling out some confounding factors, a trial's being described as double masked is a methodological value that makes it superior to open trials. However, as was pointed out in the previous sections this alleged potential advantage of double masking will not be actualized in all cases.

Furthermore, attempts to keep trials double masked have their costs. Keeping a trial double masked makes the conditions of the trial different from the conditions in routine clinical practice. In an attempted double masked placebo controlled trial, the dispenser administers an intervention that she cannot know for sure is an 'effective' (nonplacebo) treatment. The participant receives the intervention with a commensurate degree of doubt. This doubt may affect their beliefs and hence the effectiveness of the experimental or control intervention. In contrast, in routine clinical practice, the dispenser usually offers an intervention with confidence, and the patient believes that they are being treated with a nonplacebo.

It could be argued that the doubt about which treatment patients in a trial have been given does not make a difference to the estimated effects of the characteristic treatment. As long as the doubt is the same in both the test and control groups, then any reduced belief effects will 'cancel out' because they are reduced equally in the test and control group. Recall that in a placebo controlled trial the characteristic effects are calculated by subtracting the average effect in the test treatment group from the average effect in the control group. If both these groups have effects reduced by doubt, then as long as the doubt is the same in both groups, the effect estimate of the characteristic features will remain constant. An analogous argument applies to double masked active controlled trials.

The argument that the reduced belief effects 'cancel out', however, relies on the assumption of *additivity*, whereby the component features of a treatment (participant and dispenser belief, other incidental factors, and characteristic factors), simply add up to form the composite effect. If additivity does not hold, then changing the effect of

participant or dispenser belief could affect the overall effectiveness in a different way. I discuss the assumption of additivity further in chapter 8.

In sum, it seems safe to say that attempts to keep a study double masked may often fail. This may be because of the inherent difficulty in imitating an intervention's sensory qualities and side effects. The apparent methodological advantage of being double masked at the outset of a trial may therefore be frequently illusory – trials are rarely successfully double masked. If many trials are not successfully double masked, then the basis for considering double masked trials superior to open trials becomes difficult to uphold. Further, the seemingly legitimate aim of double masking studies where it seems reasonable has its costs. This does not mean that double masking should be abandoned – indeed there are many cases (trials of drugs with moderate effects and mild side effects) where double masking adds methodological value.

6.7. Conclusion

The view that double masking always adds methodological value was exposed by the Phillip's Paradox whereby, dramatically effective treatments are not supportable by double masked, and hence 'best' evidence, to be wanting. In this chapter I examined the methodological values from the more fundamental, 'scientific common sense' point of view that good evidence rules out rival hypotheses, and hence that better evidence rules out more rival hypotheses.

Double masking, where successfully executed, rules out the potential rival hypotheses (confounders) of participant expectations and dispenser attitudes. However there are three situations where these potential confounders are not actual confounders (or not significant confounders). First, there are some situations where patient and dispenser expectations do not have effects. Second, there are other cases, perhaps some 'active' controlled trials, where the expectations could be the same for both groups. Third, in cases where the experimental treatment has a dramatic effect, any expectations might have (relatively) negligible effects.

My investigation in this chapter also revealed that in practice, keeping trials successfully double masked is usually difficult and sometimes impossible. This means that being described as double masked does not necessarily mean that any methodological value has been added. In cases where it is likely that double masking will not be successful (i.e. where the side effects of the experimental treatment are common and difficult to imitate), 'open' trials lose their relative disadvantage. Indeed,

in these cases it may be better to seek other ways to control for the potentially confounding effects of expectations and attitudes. For example, by explicitly controlling for participant prejudice and by placing greater emphasis on masking the outcome assessors and statisticians. In short, although the underlying rationale for double masking – that it rules out plausible rival hypotheses – is sound, a consideration of the rationale reveals why double masking is not always a methodological virtue, and solves the Phillip's Paradox.

7. Chapter Seven. Ethics Versus Methodology: Active Controlled Trials and 'Assay Sensitivity'

It is paradoxical that there is no standard of evidence to support the standard of evidence
- (Golomb 1995)

7.1. 'Placebo' or 'Active' Control?

There are two main types of randomized trials used to provide evidence that a new intervention is more effective than a placebo[93]. In the first, the experimental treatment is compared with a 'placebo' in a placebo controlled trial, or PCT. In the second the experimental treatment is compared with an existing accepted treatment (which itself is assumed to be more effective than 'placebo') for the disorder at issue in an active controlled trial, or ACT. However, due to ethical (as well as, although these have largely been ignored, practical and methodological) constraints, PCTs are sometimes difficult to justify[94]. In cases where PCTs are problematic, the obvious solution is to employ ACTs. Yet, because ACTs have alleged methodological problems, it is often argued that they cannot replace PCTs without sacrificing the quality of the study.

The main alleged methodological problem with ACTs is often shrouded in complex detail but can be summed up quite simply. Active controlled trials, it is argued, cannot rule out the rival hypothesis that the observed effects of the experimental

---

[93] I will continue with the definitional scheme introduced in previous chapters. A 'placebo' is, on this view, a treatment whose 'characteristic' features do not have positive effects. The term 'characteristic' is borrowed from Grünbaum (1986), and is used to denote, for example, the patented chemical in a drug. Fluoxetine, for instance, would be the characteristic feature of antidepressant therapy involving Prozac. All the other features of a treatment process I will call 'non-caracteristic' (the term Grünbaum uses is 'incidental'). Factors traditionally thought of a placebogenic, such as prescription of the drug from a sympathetic physician, would be 'non-characteristic'. A legitimate placebo control must obviously control for all non-characteristic features of the treatment process. If it does not then any difference in average outcome between the placebo and experimental groups could be due to one of the uncontrolled non-characteristic features of the treatment. See chapter 4 for details on how this is, in practice, quite difficult to achieve.

[94] See appendix B to this chapter.

treatment are placebo effects without making an assumption that can only be justified by referring to information outside the trial. To justify the inference from a 'positive' result of an ACT (where the experimental treatment demonstrates that it is at least as effective as the control) we must assume that the control treatment was more effective than 'placebo'. If the established treatment control were no more effective than, or, which is possible in principle, *less* effective than placebo, then any inference from a 'positive' ACT to the claim that the experimental treatment is 'effective' will be mistaken. It must be remembered that some existing established treatments in the history of medicine have, after all, turned out to be no better than, or even worse, than 'placebos'. Bloodletting, for instance, was once widely used for many ailments, some of which are now known to be correlated with low blood pressure. A positive result of a PCT (where the experimental treatment demonstrates superiority over 'placebo'), on the other hand, allegedly allows us to make the inference to 'effectiveness' without reference to information outside the trial. I will call this argument the 'basic argument against ACTs', which is that ACTS do not rule out the rival hypothesis that the observed effect is no more than a 'placebo' effect.

The obvious response to this argument is that, if we can assume that the control treatment is more effective than placebo (and surely we usually can), then a positive result of an ACT does rule out the rival hypothesis that the experimental treatment is no more effective than a 'placebo'. It would seem, therefore, that this argument is limited in scope. Moreover, given the arguments in previous chapters that actual 'placebo' controls are illegitimate, and that even if legitimate, can have highly variable effects, it would seem that 'placebo' controls suffer from similar problems as 'active' controls.

In this chapter, I will examine the 'assay sensitivity' argument against ACTs in more detail. My first task will be to disambiguate the 3 different arguments that are often lumped together as 'assay sensitivity' arguments. I will argue that all three arguments are limited in scope, and that it is unclear whether any of them is persuasive.

7.2. The First Assay Sensitivity Argument: Debunking the Myth that ACTs Rely on a Stronger Assumption than PCTs

The first version of the basic argument against ACTs often cited in the literature is sometimes called the 'assay sensitivity argument'. According to the first definition of 'assay sensitivity' (as we will see there is a second), 'assay sensitivity' refers to the ability of a trial to detect a difference between a non-placebo and a placebo. Hence

Temple and Ellenberg state: "The ability of a study to distinguish between active [non-placebo] and inactive [placebo] treatments is termed *assay sensitivity*" (Temple and Ellenberg 2000, p. 457). The 'assay sensitivity argument' based on this definition of assay sensitivity is that PCTs but not ACTs are able to detect differences between placebos and non-placebos. This first assay sensitivity argument (as I will call it) is nothing more than the basic argument against ACTs clothed in more technical language.

I will argue that this argument is problematic in two ways. First, it only applies to the few cases where we cannot make the assumption that the experimental treatment is in fact more effective than 'placebo' and hence has a very narrow scope. It is further limited in scope by the fact that it applies mostly to 'non-inferiority' trials. Second, 'placebo' controls, like established treatment controls, are treatments in their own right and as such suffer from similar problems as established treatments. I will conclude that, outside a few cases where the control treatment's superiority to placebo is in doubt, ACTs are as good as PCTs at ruling out the hypothesis that the experimental treatment is not more effective than 'placebo'.

7.2.1. The Limited Scope of the Argument that ACTs Require the 'Control treatment assumption' for Their Interpretation

Bloodletting was once widely used for many ailments, some of which are now known to be related to low blood pressure. Even recently, there have been cases where established treatments thought to be 'effective' turned out to be worse than 'placebo'. Recall from chapter 2 that antiarrhythmic drugs such as encainide and flecanide were widely used as preventatives of ventricular ectopic beats until exposed by a PCT to *increase* mortality from heart attack relative to 'placebo'. If, in the worst case, the established treatment was much worse than 'placebo', and the experimental intervention was only slightly superior to existing treatment, then the experimental intervention, in spite of being superior to the established treatment, could still be *worse* than 'placebo'. The worry about ACTs is, then, a worry about the extent to which existing treatments are, in fact, 'effective', i.e. superior to 'placebo'.

Another way to express this general worry is to note that we have to make the 'control treatment assumption' to interpret the results of ACTs. The 'control treatment assumption' is the assumption that, because the control treatment reliably demonstrated effectiveness (usually, but not necessarily, relative to 'placebo') in the past, that we can

assume it performed more effectively than 'placebo' in the current non-inferiority ACT. In Anderson's words:

> conclusions of effectiveness in non-inferiority trials are valid if and only if the historical assumption that the active control is an effective drug is justified appropriately (i.e. by appeal to external information concerning the past performance of the drug in question) (Anderson 2006, p. 69)

The problem with justifying the control treatment assumption is usually discussed with reference to *historically controlled trials*. In these studies, a new intervention is introduced experimentally, and its effects are compared with a historical control, namely past effects of standard intervention or no intervention on allegedly similar patients. The problems with historical controls are rather obvious: we cannot generally assume that an intervention which has been successful in the past, with patients in the past, will be successful in the present with present patients. Several factors relevant to the apparent effects of the intervention change over time. These include environmental, patient, and disease characteristics, ancillary care, the type of patient that presents themselves or is selected for a trial or particular treatment (this is known as 'selection bias') or even the possibly altered virulence of the disease. Each potentially relevant difference between past and current effectiveness of control presents an alternative hypothesis for the outcome of the study *other than* the effectiveness or lack thereof of the test treatment. To rule out this set of alternative hypotheses, the control treatment assumption must be justified.[95]

Simply put, to justify the control treatment assumption is to have strong evidence to believe that the control treatment is, in fact, more effective than 'placebo'. The assumption is required to underwrite the inference from a 'positive' result of a non-inferiority ACT to the conclusion that the experimental treatment was effective, we require:

---

[95] To be sure, even where the control treatment assumption cannot be made, known differences can be adjusted for in the statistical analysis. Then, there are cases where the historical control assumption can safely be made.

> Clearly, the method can be effective in certain circumstances, namely if the natural history of the disease is so consistent and well documented that we can be sure that we are comparing like with like. If, for instance, in the past a disease has invariably and rapidly led to death, there can be no possible need for controls to prove a change in the fatality rate" (Hill and Hill 1991`, p.217).

> information external to the trial (the information about past 'placebo'- controlled studies of the active control) to interpret the results. In this respect, the ACET [non-inferiority ACT] is similar to a historically controlled trial" (Temple and Ellenberg 2000`, p. 457)

When viewed this way, it is quite clear that the argument against ACTs that they require the control treatment assumption is limited to cases where we lack good evidence to believe that the control treatments are, in fact, more effective than 'placebo'.

The fact that there is a good deal of debate about what counts as strong evidence notwithstanding, there is no doubt that appendectomies for acute appendicitis, antibiotics for bacterial pneumonia, the Heimlich manoeuvre to remove an obstruction in the airways, morphine to kill pain, and many other treatments are undoubtedly effective. Even Temple and Ellenberg admit as much: "in some settings, such as highly responsive cancers, most infectious diseases, and some cardiovascular conditions, ACETs [non-inferiority trials] can and do provide a valid and reliable basis for evaluating new treatments" (Temple and Ellenberg 2000, p. 457). Or, a few pages later, "Active control non-inferiority trials can be informative and have been used successfully and appropriately in many therapeutic areas in which assay sensitivity is not in doubt" (Temple and Ellenberg 2000, p. 459). Yet in other places they seem to suggest that the worry with ACTs is widespread: "In *many* cases, ... [the] assumption of assay sensitivity cannot be made" (Temple and Ellenberg 2000`, p.457`, italics added). Or, at the end of their paper, they claim that non-inferiority ACTs are

> *often* uninformative. They can neither demonstrate the effectiveness of a new agent nor provide a valid comparison to control therapy unless assay sensitivity can be assured, which *often* cannot be accomplished" (Temple and Ellenberg 2000`, p.457`, italics added).

I will not delve into a debate about what the precise proportion of established treatments have doubtful effectiveness, but simply note that there are many treatments for which the control treatment assumption can in fact justifiably be made and that Temple and Ellenberg may have exaggerated its importance.

It is also important to note that the argument against ACTs based on the claim that they require the control treatment assumption is obviously of lesser importance if we require that the experimental treatment demonstrate superiority rather than mere 'non-inferiority' to the control. Briefly (more will follow), a 'non-inferiority' trial, is a trial that is designed to detect whether the experimental intervention is of equal (to within some 'margin of equivalence'), or greater effectiveness than the control

treatment. 'Superiority' trials (PCTs are superiority trials), on the other hand, are designed to detect whether the experimental treatment is superior to the control. If the experimental treatment demonstrates superiority, then we have grounds to believe that it is strictly superior (i.e. neither equal nor inferior).

Even if the existing treatment were no better than 'placebo', provided it were no worse, then demonstrating strict superiority is good evidence that the experimental intervention is a positively effective non-'placebo'. This means that the control treatment assumption required in the case of a superiority ACT is simply that the control treatment is at least as effective as 'placebo'. In a non-inferiority trial, on the other hand, the control treatment assumption required is that the control treatment is more effective than 'placebo', which is obviously stronger than the assumption required in superiority ACTs. Since the assumption required for its interpretation is stronger, non-inferiority ACTs are more susceptible to this assay sensitivity argument than superiority PCTs. Implicitly recognizing this fact, it is quite clear that the target of the 'assay sensitivity' and other arguments against ACTs are non-inferiority ACTs and not superiority ACTs. For example, while Temple and claim:

> A well-designed study that shows superiority of a treatment to a control (placebo or active therapy) provides strong evidence of the effectiveness of the new treatment, limited only by the statistical uncertainty of the result. (Temple and Ellenberg 2000`, p.456).

Other authors also make it clear that the problem with ACTs is limited to non-inferiority ACTs. For instance, Hwang and Morikawa recommend that we "[s]how superiority of the test drug to the low-dose standard treatment" (Hwang and Morikawa 1999) to avoid the allegedly inherent problems with ACTs. In short, this 'assay sensitivity argument' is not aimed at superiority ACTs.

These authors, however, might be conceding too much: the control treatment could, after all, be worse than 'placebo'. Since, as noted, the probability of an established treatment being less effective than the control, at least for the target disorder, is far lower than the chances of an existing treatment control being no more effective than 'placebo'. Still, the basic argument against ACTs (as well as the assay sensitivity argument now being considered) is certainly stronger when taking non-inferiority ACTs as the target. Most treatments, even if no more effective than placebo, are probably not harmful. With this in mind, one might immediately demand a justification for non-inferiority trials. Before arguing that this assay sensitivity is

unconvincing even for the limited number of cases to which it applies, I will argue that non-inferiority ACTs are not justified as often as is sometimes claimed.

7.2.2. Questioning the Need for Non-Inferiority Trials: Pulling the Rug from Under the 'Assay Sensitivity' Argument

Non-inferiority trials are not as widely practised as superiority trials. It has been estimated that the number of non-inferiority trials is less than 2% of all randomized trials (Piaggio et al. 2006, p. 1155). However, there is evidence that most trials that claim equivalence (and hence can also claim non-inferiority) are based on superiority trials where the null hypothesis of equivalence was not rejected (Greene, Concato, and Feinstein 2000). The grounds for claiming equality or non-inferiority is, on the basis of having failed to reject a conventional hypothesis is, as has been pointed out, flawed. In brief, although non-inferiority trials may not actually be widely practiced, there are many cases where non-inferiority is claimed without a proper non-inferiority test being performed. If non-inferiority trials were employed in all cases where non-inferiority was concluded, a much higher proportion of clinical trials would be non-inferiority trials. It remains to be seen, however, to what extent the use non-inferiority trials is justified (see below).

The methodology of non-inferiority and equivalence trials, which was initially given specific attention beginning in the late 1970s (Dunnett and Gent 1977; Blackwelder 1982), has received more attention recently. For example, the widely supported CONSORT Statement[96], originally published in 1996 (Begg et al. 1996) was expanded in 2006 to include a separate checklist for the reporting of non-inferiority and equivalence trials (Piaggio et al. 2006). Stephen Senn has also dedicated a completed chapter to ACTs in the second edition of his book (Senn 2007a, chapter 15).

The rationale for non-inferiority trials is that new treatments can allegedly represent real advances without being more effective than existing treatments. Temple and Ellenberg of the United States Food and Drug Administration, and who also helped

---

[96] The CONSORT Statement contains a checklist of 21 items that are recommended for use by the authors, journal editors, and peer reviewers to improve the standards of reporting of trial quality. The Statement encourages authors to report, for instance, whether the trial was randomized, or masked.

draft the ICH E10 document, argue that: "it is critical to recognize that a new treatment might represent a major advance without being more effective than alternatives" (Ellenberg and Temple 2000`, p. 469). Similar statements have been made by others:

> While one might strive for new ... treatment to be superior to existing treatment, it is more realistic to hope for an equally good performance ... and to focus on other possible benefits of the new treatment, such as: side-effects, cost, ease of administration (Lesaffre et al. 2001`, p. 898).

The prominent medical statistician Stephen Senn (Senn 2005), as well as the revised CONSORT statement (Moher, Schulz, and Altman 2001) provide specific reasons why we might want to have several treatments for the same ailment, none of which is necessarily more effective than the next:

1. "The first is that new drug may have advantages in terms of tolerability" (Senn 2005). That is, it may have a better side effect profile, or it might be more tolerable for certain patients: "many people have an aspirin allergy" (Senn 2005).
2. The new treatment could be more available, i.e. cheaper or less invasive (Piaggio et al. 2006, p.1153).
3. The new treatment could be easier to administer, "for instance one daily dose rather than 2 doses" (Piaggio et al. 2006).
4. "introduction of further equivalent therapies before patent expiry of an innovator in the class may permit price competition to the advantage of reimbursors (Although such competition is probably not particularly effective)" (Senn 2005).

The rationale for promoting new treatments that are not necessarily more effective than existing treatments appears convincing. If the existing treatment has severe side effects, then the new treatment, even if no more effective, would be an advance if its side effects were mild. Or, if the existing treatment costs millions, then a new treatment, even if no more effective, would correctly regarded as an advance if it cost only pennies.

However before a new treatment that is no more effective than its predecessor(s) is approved for marketing, the rationale requires more serious thought than is sometimes given. I will not examine each of the alleged reasons in any detail but rather make a few general remarks.

First note is that the alleged reasons for promoting non-inferior treatments also support the introduction of a superior treatment. A new treatment could be more effective as well as have a better side-effect profile and be cheaper. Surely it would be better if the new treatments were better than the existing treatment rather than simply non-inferior. There is no reason why, to be prepared in the event of intolerability of the new treatment, some older, less effective treatments could be kept in case the new one is not tolerable for some patients.

Then, the claim that a new treatment could represent an advance without being more effective because it has a better side effect profile is not as uncontroversial as it sounds. Existing treatments have usually been around for longer, so there will be more extensive and accurate data about their rare and long-term side effects. The new treatment, on the other hand, will have a less extensive and accurate side effect profile simply because it typically will not have been around for as long. Hence the comparison based on side effects will often be unbalanced, a point which is often unconsidered.

Next, the reasons for having more than one treatment that is no more effective than an existing treatment only go so far. Even if useful, the reasons become less convincing as the number of non-inferior treatments increases to several. For example, there are currently over 6 SSRI antidepressants on the market (fluoxetine, sertraline, citalopram oxalate, escitalopram oxalate, fluvoxamine maleate, and paroxetine). In addition, there are other pharmaceutical antidepressants (tricyclics, monoamine oxidase inhibitors (MAOis), serotonin-norepinephrine reuptake inhibitors (SNRIs), noradrenergic and specific serotonergic antidepressants (NASSAs), norepinephrine (noradrenaline) reuptake inhibitors (NRIs), and Norepinephrine-dopamine reuptake inhibitors) of which there are over a 12 dozen in all. There are also several other treatments used to treat depression, such as St. John's Wort, Cognitive Behaviour Therapy (CBT), exercise, and self-help. None of these treatments have demonstrated consistent superiority to the other in clinical trials, although the administration of some (e.g. exercise) is admittedly very different from others. Even if it were useful to have a few of these treatments on hand, it is questionable whether we need so many. The rationale for conducting non-inferiority trials surely becomes weaker as more and more treatments become available.

The case for non-inferior treatments of the same class as the existing treatment is also weaker than a new non-inferior treatment of a different class. Drugs of the same class, because they have similar mechanisms of action, are also likely to have similar

side effects and administration regime. The case for non-inferior treatments of the same class (or even of different classes), also becomes weaker as the number of treatments in that class increases. Indeed the use of non-inferiority trials has been condemned on these grounds: "Some noninferiority trials have been criticized for merely studying a new marketable product ('me too' drugs) (Piaggio et al. 2006', p.1152).

More importantly, if the alleged justification for a non-inferiority trial is that it has fewer side effects, it would seem that a superiority test of the relevant side effects would be the correct choice of treatment. It is possible, of course, to run a superiority test for the side effects of interest and a non-inferiority test for the main outcome at the same time.

To take stock, the rationale for using non-inferiority trials is acceptable when we have good reason to believe that another treatment for the same ailment is likely to provide a real benefit. Real benefits can include different or more favourable side effect profiles, different interactions with other medications, and even cheaper cost. The case, however, for more than two or three drugs of the same class, which are likely to have similar characteristics, is more difficult to make.

I will now leave aside the issue of whether non-inferiority ACTs are justified, and argue that 'placebo' controlled trials require a similar control treatment assumption as 'active' controlled trials.

7.2.3. Placebo Controls as Treatments that Suffer from the same Problems as established treatment controls

If the 'placebo' control is illegitimate, then, like active controls, they could be more effective than legitimate ('real') 'placebo' controls, or less effective than legitimate 'placebo' controls (see chapter 5 for details). At least in principle, 'placebo' controls could even be harmful. Recall from chapter five that olive oil had been used in the 'placebo' controls for cholesterol lowering agents before it was known that olive oil reduced cholesterol. These 'placebos' could well have been more effective than legitimate 'placebo' controls and led to false claims about the ineffectiveness of the cholesterol lowering agents[97]. Moreover 'placebo' pills will sometimes contain

[97] Golomb (1995) states: "several early papers exploring the use of cholesterol-lowering agents to curb heart disease did in fact name the placebos used: olive oil in one case, and corn oil in another. Mono- and poly-unsaturates such as olive oil and corn oil are now widely known to

additional agents to make them taste and smell like the test treatment. It is difficult to know in advance and without further investigation whether these additional agents will have positive, or even negative effects on the target disorder. In one example, an ingredient (cellulose acetate phthalate) typically used to coat pills was found to have positive effects for curing several sexually transmitted diseases, including herpes (Gyotoku, Aurelian, and Neurath 1999; BBC News 1999). Although in the examples cited the 'placebo' control had additional *positive* effects, there is no reason why this should necessarily be the case.

In another example discussed in much more detail in chapter 6, Kirsch found evidence that the superiority of SSRIs over 'placebo' could be explained by unmasking of the studies rather than the actual effectiveness over the 'placebo' effect of the drug itself (Kirsch and Sapirstein 1998). This could have happened if the participants taking the 'placebo' knew they were taking the placebo and hence had lower expectations, while the participants taking the experimental treatment knew they were taking the 'real' treatment and had higher expectations. The different expectations, and not the characteristic features of the treatment, then, could have explained the difference between experimental and 'placebo' treatments.

The fact that illegitimate 'placebos' may be common is sufficient by itself to warrant a revision of the view that ACTs require a stronger assumption about the effectiveness of the control treatment than PCTs. This fact may have been overlooked because of a failure to consider more carefully what counts as a legitimate 'placebo' control. Having done this (see chapter five), however, it becomes clear that both PCTs and ACTs face similar problems. The fact that even the same 'placebo' could have very different effects in different circumstances makes the case for requiring a strong assumption to interpret PCTs even stronger.

Whether a 'placebo' pill is blue or red, whether a 'placebo' is injected or taken orally, and whether it is presented as being a brand name[98] can alter the effects of the 'placebo' (Anderson 2006`, p.73). Citing evidence from an earlier study (Moerman

---

decrease low-density lipoproteins, so that with hindsight these agents may not have been inert with respect to the outcome studied. Indeed, it was noted in one such study that the rate of cardiac mortality was lower in the placebo group than expected".

[98] (de Craen et al. 2000) (Nagao et al. 1968; Huskisson 1974; de Craen et al. 1996) (Branthwaite and Cooper 1981)

1983), Anderson notes that the 'placebo' response rate, like the effectiveness of standard treatment controls, can vary greatly even in the absence of apparent differences of colour and modality.

> In a meta-analysis of 30 PCTs of cimetidine [a treatment for duodenal ulcers], Moerman found that about half the trials declared that cimetidine was superior to placebo ... the other half declared that cimetidine was no better than placebo. The different conclusions were not accounted for by the response rate to cimetidine; this was strikingly constant across all trials (70-75%). Rather, it was the placebo response rate, which varied incredibly, from a low of 10% to a high of 80% [success, defined as endoscopically observed healed ulcer craters], that accounted for the difference (Anderson 2006`, p.74)[99].

In the trials where the 'placebo' response rate was high, cimetidine appeared to be ineffective; where it was low, cimetidine appeared to be effective. If it is true that 'placebo' response rates vary greatly, then it is true that PCTs require something analogous to the control treatment assumption to rule out the hypothesis that the experimental treatment is no more effective than 'placebo'.[100] Given the evidence from Moerman, Anderson concludes that an outcome of a PCT that seems to show superiority provides, when taken on its own, no firmer grounds for concluding effectiveness than does a non-inferiority outcome in an ACT, does not by itself provide evidence that the test treatment was effective.

In fact the evidence for the variability of 'placebo' controls is far stronger in the cimetidine study than even Anderson suggests. The variability of the initial 31 (not 30 as Anderson states) cimetidine studies was from 10% - 90% (not 10% - 80% as Anderson states). "In the 31 studies of ulcer treatment [cimetidine] reviewed here, 'placebo' effectiveness ranged from 10% to 90% with a mean of 45.6% and a standard

---

[99] The Moerman study seems to contradict the Hróbjartsson and Gøtzsche analyses cited last chapter. However, the Moerman study was very different since it did not compare the effects in the 'placebo' group with the effects of 'no treatment'. Hence nothing can be concluded about the magnitude of expectation effects from the Moerman study. Moreover, Hróbjartsson and Gøtzsche do not claim that there are no placebo effects *simpliciter*, but merely that their systematic review revealed insignificant ($\neq$ non-existent) placebo effects outside trials of treatments with subjective and continuous outcomes (which accounted for more than half the trials in their review).

[100] Incidentally, the variability of the placebo response calls into question the claim that PCTs provide a measure of absolute effect size (see next chapter).

deviation of 18.8%" (Moerman 1983`, p.13). Further, Moerman's initial study was expanded in 2000. In the more extensive study, the number of cimetidine trials examined increased from 31 to 74, and this was coupled with results from 45 ranitidine (a different, but similar intervention for duodenal ulcer) trials (Moerman 2000`, p.58). The follow-up study came to the same conclusion as the first: placebo rates differ widely – indeed in the more extensive study the 'placebo' response rate ranged even more widely, from 0 to 100%.

Moerman also did his best to rule out obvious sources of 'placebo' variability. He noted no statistically significant effect of participant age or gender, length of study, or frequency of treatment (Moerman 2000`, p.62). There were, however, national differences; the 'placebo' response rate was less than half the average in Brazil and twice the average in Germany. The national differences, however, were almost exactly reversed in placebo controlled trials on interventions for hypertension and generalized anxiety disorder. Moerman infers that "Placebo effects seem to be highly variable regardless of the axis on which you examine them" (Moerman 2000`, p.64). He concludes:

> [t]his finding is highly counterintuitive for medical researchers who usually consider placebo effects to be constant – sort of like "noise" in the system – and orthogonal, unrelated to medical effects. Both of these notions are clearly incorrect (Moerman 2000`, p. 58).

Even critics of Moerman's study such as Stephen Senn admit that the placebo response rate is highly variable, and that the response rate determines whether cimetidine was judged 'effective':

> the response rate under placebo in a given trial is a good predictor of the response under cimetidine. This is a very reasonable conclusion and may be expected to apply generally for many different treatments and conditions (Senn 2006`, p. 1).

If the (variable) 'placebo' response rate is a good predictor of the apparent effectiveness of the test treatment, then it is unclear how PCTs rely less strongly on an assumption about their effectiveness than do ACTs. To infer from positive results of PCTs to effectiveness of the test treatment we must infer that the 'placebo' control was not (for some reason) less effective than what it should have been.

Or, it could be objected that the single study Anderson cites of the variability of the 'placebo' effect is hardly sufficient to conclude that *in general* 'placebo' controls can vary much like standard treatment controls. Temple and Ellenberg could counter that the variability of certain interventions 'known' to be effective has been

demonstrated still more frequently – they note 11 cases (see below). However, the evidence for the variability of active treatments is not as strong as Temple and Ellenberg imply (see next section). Until extensive empirical work is done, it can, it seems, be argued that the control treatment assumption is equally justified (or unjustified) for both PCTs and ACTs.

In sum, the number of treatments whose effectiveness can be doubted is limited. Furthermore, the treatments used as placebo controls in trials can be both more, or less effective than the 'true' placebo. Therefore, they are best viewed as treatments in their own right that suffer from all the problems as existing established treatment controls. More specifically, PCTs cannot, without making assumptions about the 'placebo' control, rule out the hypothesis that the experimental intervention is no more effective than 'placebo'.

It has been objected, however, that even if we can make the control treatment assumption, that ACTs still face the basic problem. I will now consider this objection.

7.3. Lack of Sensitivity to Effects or Ineffectiveness?

It has been argued that, even when we can make the assumption that the control treatment is more effective than 'placebo', positive results of ACTs still do not rule out the rival hypothesis that the experimental treatment is no more effective than placebo. This is because, it is alleged, some treatments that we know are more effective than placebos (i.e. for which the control treatment assumption is justified) simply do not allow us to detect their effects reliably:

> the effectiveness of drugs that sometimes (or even often) fail to be proven superior to placebo is not in doubt; even if a drug is statistically significantly superior to placebo in only 50% of well-designed and well-conducted studies, that proportion will still be vastly greater than the small fraction that would be expected to occur by chance if the drugs were ineffective (Temple and Ellenberg 2000`, p. 458).

If there are such treatments, then if ACTs employ these treatments as controls, we cannot infer from a positive result to 'effectiveness' of the experimental treatment. That is, a positive result of an ACT will be ambiguous: the established treatment control, although effective in a general sense, might have failed to demonstrate its effectiveness in that particular trial.

I will call this argument the 'sensitivity to effects' argument, since the argument relies on the claim that in spite of 'undoubted' effectiveness, trials of certain treatments are simply not sensitive to their effects. If trials of certain treatments are simply

insensitive to effects (that is, if they have what I shall henceforth call 'sensitivity problems'), then there are cases where we can make the control treatment assumption yet where PCTs are more reliable than ACTs.

A 'positive' PCTs (where the experimental treatment demonstrates superiority) allegedly provides internal confirmation that the trial was sensitive to the characteristic effects of the experimental treatment. If it detected superiority to placebo, then we can be sure, to a degree limited by the statistical uncertainty of the result, that the trial was able to detect superiority to placebo A negative result, on the other hand, does not: the trial may have been insensitive to the non-placebo effects. In this case the drug could (allegedly) have a real and indubitable effect yet it simply failed to demonstrate its effectiveness in that particular trial.

The first thing to note is that a strict interpretation of the claim places the problem with trials (of certain treatments) rather than the effectiveness of the treatments themselves. Whereas the basic problem with ACTs is ontological and has to do with properties of the control treatments, this alleged problem is epistemological and has to do with trials (of certain treatments). I will examine this argument on its own terms and argue that, contrary to what Temple and Ellenberg claim, if well conducted trials fail to consistently detect effects over and above placebo effects, then it is reasonable to doubt the effectiveness of the treatments themselves rather than the trials. If the problem is with the treatments, then the 'sensitivity' argument becomes the basic problem with ACTs that not all potential control treatments are, in fact, more effective than placebos.

7.3.1. Temple and Ellenberg's Unjustified Reasons for Blaming the Tools

Replicability is a central tenet of scientific method. If the results of an experiment are not replicable, then there are two sensible interpretations. Either (a), one of the attempts to replicate was flawed (and the attempt to replicate cannot be said to have failed), or (b) the failure to replicate indicates that the original experiment cannot be trusted. I will consider each of these possibilities in turn. Temple and Ellenberg seem to reject the possibility that the reason trials do not detect the effects of certain treatments is that the trials were flawed in some way:

> One might speculate that variable results of trials of antidepressants are simply the consequence of modest effect sizes coupled with samples too small too overcome the inherent variability of the condition studied. Results, however, are consistent with effect sizes that vary greatly and unpredictably from study to study. With current knowledge, one cannot specify a

> particular study population treatment protocol, or sample size that will regularly identify active agents (Temple and Ellenberg 2000`, p.458)

Taken seriously, their statement it can practically be reduced to the absurd. Imagine a well-conducted study that involved the entire target population, say all depressed people in the world. Imagine that the trial failed to demonstrate difference between experimental treatment and 'placebo'. Surely we would want to say that this study provided strong reasons to doubt the effectiveness of the experimental treatment. Yet when they claim "one cannot specify a particular sample size ... that will regularly identify active agents", they are free to blame "current knowledge" (or rather "current ignorance") for failing to design a sufficiently sensitive trial.

Having rejected the possibility that the trials were flawed in some currently known way, Temple and Ellenberg other reasons for us to accept the idea of treatments with 'sensitivity issues':

> In the cases described, the effectiveness of drugs that sometimes (or even often) fail to be proven superior to placebo is not in doubt ... a generally small response that varies among populations, insufficient adherence to therapy or use of concomitant medication, study samples that improve spontaneously (leaving no room for drug- induced improvement) or that are unresponsive to the drug, or some other reason not yet recognized. What all these influences have in common is that they reduce or eliminate the drug-placebo difference, so that a study design and size adequate to detect a larger effect will not detect the reduced effect(Temple and Ellenberg 2000`, p. 458-9).

It is surely odd, however, to claim that the supposed problem is only identifiable by "observed failure of the trial to distinguish drug and placebo treatments". Surely it is possible that the trial was in fact sensitive to effects, but that the treatment, at least for that trial population, *was no more effective than placebo*. In fact, contrary to what they claim, the reasons listed do not provide sufficient reason to reject the hypothesis that treatments with 'sensitivity problems' are ineffective. To begin, small effects that vary across populations are not sensitivity problems *per se*, but rather a problems with the size of the trial that can be remedied with larger trials and the use of systematic reviews. For example, the effects of aspirin for reducing mortality in patients suspected of acute myocardial infarction were thought to be non-existent on the basis of evidence from small trials. But when a much larger trial was conducted, the effects, although small, were significant (Baigent et al. 1998). Because of the severity of the incidents prevented by aspirin (stroke and death), the small effects are thought to be important (Peto, Collins, and Gray 1995).

Next, insufficient adherence to therapy and concomitant medication are not sufficient explanations for why trials might not detect an effect. Randomized trials routinely employ pre-trial screening to filter out participants for various reasons including concomitant medication, and usually have run-in periods to eliminate participants who don't comply. So Temple and Ellenberg are quite wrong in this respect that the judgement that there was differential adherence to the treatment in the two groups can only be motivated *a posteriori* – that is after finding a negative outcome for a drug 'known' to have a positive effect. Certainly patients are more likely to be followed to ensure adherence in an RCT than they are in general practice. The fact that some highly artificial conditions could be created in which a treatment might demonstrate a (mild) characteristic effect does not mean that the treatment is an effective non-placebo in any accepted sense. In fact, attempts to tighten the conditions of the (many would argue already too different from routine practice) of clinical trials more might make it *overly* sensitive to effects, which is to say that it might detect a mild characteristic effects that would be inexistent in routine practice.

Then, the claim that study samples improve spontaneously leaving no room for drug-induced improvement, rather than invalidating the test, seems clearly to support the hypothesis that the treatment is ineffective. There may exist, for example, a treatment that makes symptoms of the common cold disappear within 2 weeks. However, since symptoms of the common cold almost always disappear at least as soon, the treatment will usually not have a chance to demonstrate its effects. To call such a treatment effective seems like a misuse of language. Yet when Temple and Ellenberg blame "study samples that improve spontaneously (leaving no room for drug- induced improvement) or that are unresponsive to the drug", they seem to do just this. Such a treatment may, of course, be of great benefit to those whose colds do not disappear within two weeks. However we will not know this until we have identified these patients and tested the new treatment on them in trials that were, in fact, sensitive to effects. If an intervention is incapable of acting even as a catalyst for spontaneous improvement over large and varied samples, then we certainly have reason to doubt, if not deny, the superiority of such a treatment to 'placebo'.

A similar argument applies to the claim that treatments may not consistently demonstrate effects because of 'unresponsive' participants. If participants are unresponsive, then the treatment is, for those participants, ineffective. Until we identify a group of people for which the treatment is effective (in trials that are sensitive to

effects), because of the potential side effects and costs, we certainly have grounds for doubting the effectiveness of the treatment.

Then, attributing a reduced drug-'placebo' difference to some unknown reason is not acceptable. Unless some such reason is articulated and positively tested thus becoming a known reason, we cannot justify the claim that unknown reasons will reduce the drug – 'placebo' difference. Unknown reasons, by definition, could increase or decrease the drug-'placebo' difference. This is typical of a degenerating *ad hoc* step.

In response, Temple and Ellenberg claim that the problems with the trials they list are only tentative and that the real reason they know a trial lacked assay sensitivity is the fact that it failed to detect a treatment effect:

> In each case, however, the problem [of *in*sensitivity for a particular treatment] is not identifiable a priori by examining the study; it is recognized only by the observed failure of the trial to distinguish the drug and placebo treatments (Temple and Ellenberg 2000`, p. 459).

This response, however, is circular. As an argument against the possibility that the treatments with assay sensitivity problems are in fact ineffective, it begs the question. Temple and Ellenberg could, of course, counter that they might one day articulate a hitherto unknown reason that explains why it is that an (ontologically) 'effective' treatment fails (epistemically) to reveal its effectiveness in some given trial. This suggestion, however, is untestable. The way we discover the ontological property of effectiveness is with trials. In short, the claim that we only know that the trial lacked assay sensitivity, as opposed to the alternative hypothesis that the treatment lacked 'effectiveness', is untestable.

To summarize, the reasons Temple and Ellenberg provide for blaming the trials for not detecting allegedly indubitable effects are, on the whole, wanting. This makes the alternative hypothesis, namely that the treatments themselves are no more effective than placebo, more attractive. I will now consider the argument that the treatments with 'sensitivity problems' are ineffective in more detail.

7.3.2. Temple and Ellenberg's Probabilistic Argument for Asserting the Effectiveness of Treatments with 'Sensitivity Problems'

As an argument for the view that treatments with 'sensitivity problems' are in fact effective, Temple and Ellenberg then employ probabilistic reasoning:

> In the cases described, the effectiveness of drugs that sometimes (or even often) fail to be proven superior to placebo is not in doubt; even if a drug is

> statistically significantly superior to placebo in only 50% of well-designed and well-conducted studies, that proportion will still be vastly greater than the small fraction that would be expected to occur by chance if the drugs were ineffective (Temple and Ellenberg 2000`, p. 458).

At current levels of significance, an intervention will demonstrate superiority to 'placebo' 1 in 40 times due to chance alone. Even if an intervention proved effective in merely 50% of the tests, this far outweighs what would be predicted by chance alone on the assumption of no real effect (or rather no extra effect over placebo). Indeed SSRIs, the example most extensively used by Temple and Ellenberg, actually do fail to demonstrate superiority to 'placebo' in up to 50% of cases:

> Overall, in recent experience at the U.S. Food and Drug Administration, about one third to one half of modern antidepressant trials do not distinguish a known effective drug from placebo (Laughren T. Unpublished observations (Temple and Ellenberg 2000`, p. 458).[101]

It is, however, questionable to conclude that an intervention which demonstrates superiority to 'placebo' in only 50% is in fact, effective. It could be less effective than 'placebo' in the other 50%. This means that on average (assuming that the trials were all of similar size), we can't be sure whether an intervention with sensitivity problems is effective if we follow Temple and Ellenberg's reasoning.

In their defence, however, it could be the case that the trials which fail to demonstrate an effect simply fail to demonstrate a statistically significant effect, while those that do show an effect show a dramatic effect. This possibility is relevant but is not considered by Temple and Ellenberg. Yet, it would seem that systematic reviews of SSRIs, which amalgamated the results of all sufficiently high quality RCTs could settle this issue.

---

[101] Of the 3 studies considered by the US FDA for the approval of fluoxetine (Prozac), only one clearly showed effectiveness. One study showed no difference between fluoxetine and placebo (Rickels et al. 1985). Protocol 27, a six-centered study, showed Prozac to be inferior to imipramine but superior to placebo. Some of the data in Protocol 27 was called into question and the study was omitted at the request of the FDA. With the data removed, the effectiveness of Prozac was unclear. A small study (total of 84 patients at the outset of the trial – only 11 people in the Prozac group completed their course), showed a mild benefit of Prozac over placebo (Fabre and Putman 1987)..

In fact, systematic reviews of SSRIs have had ambiguous results. Although the Cochrane Collaboration systematic review concludes that there are small differences between active 'placebo' and SSRIs (Moncrieff, Wessely, and Hardy 2004), other systematic reviews have denied that there are any significant benefits of SSRIs over 'placebo' (Kirsch and Moore 2002). I will not evaluate the systematic reviews here but simply note that is a controversy over whether overall, SSRIs have effects over and above 'placebo' effects, and that it is therefore unjustified to claim that the effectiveness of SSRIs is beyond a doubt.

Furthermore, the probabilistic argument only works if we ignore the possibility of systematic bias. I will not review the evidence that publication bias (De Angelis et al. 2004) funding source (Leopold et al. 2003; Yaphe et al. 2001), and unmasking (see chapter 6) can systematically exaggerate both the direction and size of the apparent effect. What is important to note for present purposes is that it is fair to say that these and other undetected biases will generally lead to an overestimation of the test treatment effect size. Given the small effects (relative to 'placebo') of SSRIs, these small biases could quite easily have tipped the scales in many – indeed even up to 50% - of cases. It is also relevant that systematic reviews and meta-analyses run into the same problems with publication and other systematic biases. Temple and Ellenberg are silent on this important point.

It could be objected that even if the hypothesis that SSRIs are ineffective needs to be taken more seriously, the other interventions with apparent sensitivity problems are in fact undoubtedly effective. This is difficult to judge from the paper. As Anderson notes, "Temple and Ellenberg provide empirical support for no more than four of the classes cited as drugs with assay sensitivity problems" (Anderson 2006, p. 70). Of the 11 interventions listed by Temple and Ellenberg [102], they provide evidence for only 4 (SSRIs, analgesics, postinfarction beta-blockers, and antiemetics – at any rate one antiemetic, namely ondansetron). In short, the scope of the sensitivity argument, even if accepted on its own terms, seems to be very limited.

---

[102] SSRIs, analgesics, anxiolytics, antihypertensives, hypnotics, antianginal agents, angiotensin-converting enzyme inhibitors for heart failure, postinfarction beta-blockers, antihistamines, nonsteroidal asthma prophylaxis, motility-modifying drugs for gastroesophageal reflux disease, "and many other effective agents" (Temple and Ellenberg 2000`, p. 458).

Moreover, it is unclear how the evidence Temple and Ellenberg cite supports their claim that "many" undoubtedly effective treatments have sensivity problems. They cite a single 1994 paper to support the claim that analgesics have assay sensitivity problems (Max 1994), but it is well-known that analgesics are not all effective for all types of pain management (Motamed et al. 2000). Thus in the case of analgesics, the variable results of PCTs may have to do with the fact that some analgesics are effective for some types of pain, while others are not. Within certain defined classes of disorders, certain analgesics may not have 'sensitivity problems' at all. Also, Temple and Ellenberg cite a single 1985 paper to support the claim that beta-blockers have assay sensitivity problems (Yusuf et al. 1985). But the authors of this paper encourage the use of large trials and systematic reviews to detect the intervention effect. "Although most trials are too small to be individually reliable, this defect of size may be rectified by an overview of many trials" (Yusuf et al. 1985). This suggests that the explanation for the variable results of beta-blockers is that the effect size is small. Evidence that antiemetics have assay insensitivity problems is given by a single paper (Tramer et al. 1998). Although the paper supports Temple and Ellenberg's argument that assay sensitivity may be a problem for ACTs and not PCTs, they don't deny that the efficacy of antiemetics can be established in a large enough trial or systematic review, which means that the problem may be a small effect size. In short, it is unclear that the four classes of interventions that Temple and Ellenberg discuss in greater detail have assay sensitivity problems rather than 'effect size' problems. The effect size problem is easily remedied by using a systematic review, or by conducting larger individual trials.

More relevantly to the debate over whether non-inferiority ACTs are inferior, PCTs will be equally bad evaluators of new treatments that have 'sensitivity issues'. For example, if a new SSRI is developed, and it has 'sensitivity' issues like other SSRIs (a reasonable assumption), then we cannot expect a PCT to detect its effectiveness even if it is 'undoubtedly effective'. In that particular assay it might not show up as effective. In this case the new, supposedly 'undoubtedly effective' SSRI might fail to be approved for marketing in spite of its effectiveness. Of course if it did appear effective in the PCT then sensitivity can allegedly be assumed, but if the first several trials of this treatment do not demonstrate superiority to 'placebo', the treatment could be dropped before its alleged superiority to 'placebo' was detected.

Furthermore, given the high variability of placebo response cited in the previous section, it is certainly possible that some placebos have 'sensitivity problems' as well.

In fact, the reason SSRIs and other treatments with 'assay sensitivity' issues fail to demonstrate consistent superiority to placebo could be due to the variable effectiveness of the 'placebo' controls themselves. In order to attribute the variability to the standard treatment control and not the 'placebo', further empirical research must be conducted.

It is also relevant that the sensitivity argument against ACTs, as it is presented, only applies to non-inferiority ACTs. Provided that treatments with 'sensitivity problems' do not come up as less effective than 'placebo', superiority to these treatments is strong evidence that the experimental treatment is a positively effective non-placebo.

One final remark: one might wonder why three-armed trials, with both an existing treatment and 'placebo' control, are not used. Temple and Ellenberg suggest that the three-armed solution is ideal:

> A three-arm study (new drug, placebo, and active control) is optimal because it can 1) assess assay sensitivity and, if assay sensitivity is confirmed, 2) measure the effect of the new drug and 3) compare the effects of the two active treatments (Temple and Ellenberg 2000`, p.456).

However the three-armed solution is only ideal if the apparent benefits of PCTs over ACTs are real. I hope to have shown that superiority based on assay sensitivity is illusory. Further, the three-armed solution does not address the ethical problem with PCTs, which is based on administering 'placebo' when treatments are available that are believed to be 'effective'. Also, three-armed trials are practically more difficult because they require more participants. Of course, if the ethical or practical issues do not arise, then including a 'placebo' group wouldn't do any harm.

To sum up, if well-conducted trials fail to detect characteristic effects of treatments regularly there are two plausible explanations. First, the trials could be flawed, and second, the treatments might have no effects. On the first interpretation, the trials that failed to detect an effect were flawed, and can hence be omitted from consideration. This would leave us with the set of trials in which the treatments were positively effective, and lead us to conclude that the treatment was, in fact, effective. Yet Temple and Ellenberg fail to provide us with adequate reasons for rejecting any of the trials. Contrary to what they claim, if good trials repeatedly fail to detect effects, there is good reason to suspect the superiority to 'placebo' of the treatments. Temple and Ellenberg's probabilistic support for the claim that there are treatments whose effectiveness is beyond a doubt yet whose superiority to 'placebo' has not been demonstrated predictably in clinical trials also fails. The probabilistic argument does not

take systematic bias, or the trials where the treatments fail to demonstrate superiority to 'placebo' into account. If the problem with 'sensitivity' is with the effectiveness of the control treatments, then the 'sensitivity argument' against ACTs can simply be viewed as a warning that not all potential control treatments are effective. As such, it is merely a reiteration of the basic problem with ACTs that has already been dealt with above.

I will now consider the second 'assay sensitivity' against ACTs in more detail.

7.4. The Second 'Assay Sensitivity' Argument: 'Superiority' and 'Non-Inferiority' Trials are Equally Assay Sensitive

Up to now, I have taken the definition of 'assay sensitivity' to be the ability of a trial to detect the difference between a placebo and a non-placebo. The term 'assay sensitivity', however, has also been used in a very different way. The International Conference on Harmonization E10 document, produced by the regulatory bodies of the United States, European Union, and Japan, states that: "*Assay sensitivity* is a property of a clinical trial defined as the ability to distinguish an effective treatment from a less effective or ineffective treatment (ICH 2000, p. 7). Hwang and Morikawa (1999) offer a similar definition: "Assay sensitivity refers to the ability of a specific trial to detect a difference between treatments, if one exists" (p.1208). The second 'assay sensitivity argument' is then the argument that PCTs but not ACTs are able to detect a difference between more and less effective treatments.

The second assay sensitivity argument is basically that the statistical tests used in ACTs (or at any rate *certain* ACTs, namely *non-inferiority* ACTs) are inherently incapable of detecting differences while the statistical tests used in PCTs are. In this section I will argue that this version of the 'assay sensitivity argument', although often shrouded in complex statistical detail, is either (a) merely a version of the argument that the control treatment assumption is required by ACTs but not PCTs, or (b) unacceptble.

To understand the argument, the difference between the statistical tests used in non-inferiority and superiority trials must be explained.

7.4.1. Superiority and Non-Inferiority Trials: Some Terminology

Fisherian significance tests purport to disprove a hypothesis (the 'null hypothesis' or 'null' for short). In tests using the Neyman-Pearson framework, which includes most medical trials, there is also an alternative hypothesis. If the null

hypothesis is rejected, then the alternative is 'accepted', although, as cannot be over-emphasized, rejection of the null does not strictly speaking imply anything according to the orthodox view. Like the logic of Popper's falsification, one interpretation of the logic of significance tests[103] do not permit us to confirm anything; rather, they permit is to falsify or 'reject' the null hypothesis. This means that we set up the statistical test so that, in a sense, we rule out the opposite of what we would like to accept[104].

For example, in a 'conventional' hypothesis test such as those used in placebo controlled trials, we would like to rule out the possibility that the experimental treatment and placebo are of equal effectiveness. Hence the null would be that of no difference between experimental treatment and placebo. A 'positive' result of a PCT will then be a PCT where the null hypothesis of equality was rejected, which would mean that there is either a positive or negative difference between experimental and control treatments.

In Neyman-Pearson test, the nature of the alternative will help define the test. Since, in a placebo controlled trial, we are seeking a positive difference, the alternative hypothesis will be that the experimental treatment is superior to the placebo. These tests will be 'one-tailed' or 'one-sided' since we look for a difference in only one (the positive) direction. For obvious reasons these tests are called 'superiority trials'. In a superiority trial, then, the null hypothesis is the hypothesis that the experimental treatment was of equal or lesser effectiveness than the control treatment[105].

---

[103] Gillies (1990) suggests a way for what he calls his 'neo-Popperian', as well as classical significance testing can be combined with confirmation theory (Gillies 1998).

[104] In practice, however, rejection of the null hypothesis is taken to mean acceptance of the alternative. Indeed if we were not allowed to take liberties with the interpretation of classical hypothesis tests, and, in practice, 'accept' the alternative hypothesis when the null was rejected, then they would be difficult to interpret in such a way as to guide action. For instance, if the null hypothesis that there is no difference between a treatment and a placebo was consistently ruled out in favour of the alternative of superiority, we (surely justifiably) would infer that the experimental treatment was in fact effective. I will therefore allow myself to speak loosely and assert that when the null is rejected, that it is acceptable to 'accept' the alternative.

[105] Of course, to run the test, a determinate probability distribution under the null and alternative hypotheses (a probability distribution given that the null or alternative hypotheses are true) must be specified. To say that the null is of no difference *or* inferiority, does not allow us to specify a determinate probability distribution is therefore, in a sense, strictly speaking false. In fact the

A 'negative' result of a PCT, i.e. where the null of no difference is not rejected, does not provide evidence for the truth of the null hypothesis of no difference, or equivalence[106]. In spite of this, equivalence (or, more accurately, rough equivalence) is often claimed after failure to reject the null hypothesis in a conventional test (Piaggio et al. 2006, p.1155).

In order to provide evidence for (rough) 'equivalence' using the logic of classical hypothesis testing, in, we would have to reject *two* null hypotheses: the null hypothesis that the experimental treatment was inferior (by more than some minimum amount delta[107]) than the control treatment, and the null that it was more effective (by at least some minimum 'equivalence margin' delta) than the control treatment. The role of the null hypothesis is, in a sense, reversed when comparing equivalence with 'conventional' tests. In the former, the null hypotheses are those of difference, while in the latter it is of no difference.

However we are not generally interested in rough equivalence, but whether the experimental treatment is more effective than the control treatment. In some cases, it is argued (I consider this argument below), it is sufficient for the experimental treatment to be of equal or superior effectiveness. To find out whether the experimental treatment was at least as effective as the control treatment, we would have to rule out strict inferiority. That is, we would want to rule out the possibility that the experimental treatment was less effective, by at least some minimum amount 'delta', than the control treatment. The alternative hypothesis would then be that the experimental treatment was of equal or greater effectiveness.

---

null hypothesis of no difference is used as a boundary. If a positive difference is found, then the null of equality or inferiority is said to be ruled out.

[106] There is, to be sure, a large body of literature that calls into question the logic of classical significance tests, what 'rejecting' the null hypothesis means, whether *P*-values are meaningful, and if so to what extent the value warrants a rejection of the null, and if it does, whether rejection of the null warrants acceptance of the alternative hypothesis (Howson and Urbach 1993; Lindley 1982, 1993). I will not examine this literature here, but take classical hypothesis tests as given since they are still almost universally used in medical trials.

[107] The correct choice of delta is usually chosen as "the smallest value that would be a clinically important effect" (Piaggio et al. 2006). However, the determination of the "smallest clinically important effect" is open to a certain amount of interpretation. See appendix B for some discussion and references.

A non-inferiority test is therefore, in a sense, half of an equivalence test. Rather than setting up two null hypotheses as we would in an equivalence test, i.e. those of inferiority and superiority, only the hypothesis that the experimental treatment is less effective than the control (by some minimum amount 'delta'). The alternative hypothesis is then that the experimental intervention is of equal or greater effectiveness than the control intervention. A positive result of a non-inferiority trial implies that we accept the alternative hypothesis.

The relationship between Type I and Type II in both superiority and non-inferiority trials must also be understood in order to follow the subsequent discussion. A Type I error, or a 'false positive', is the error of rejecting a true null hypothesis (and instead 'accepting' a false alternative hypothesis). A Type II error, or 'false negative', is the mistake of failing to reject a false null hypothesis (and 'rejecting' a true alternative hypothesis). Both Type I and Type II errors can be controlled for and specified in advance of a trial. When the Type I error rate, or 'significance level' is sufficiently low, generally 1% or 5%, then rejection of the null hypothesis is said to be justified[108].

In a superiority trial, a Type I error is the mistake of rejecting no difference (and 'accepting' that the experimental treatment is superior) when there is a difference. If the Type I error rate, or α, in a superiority trial is sufficiently low, then the trial will probably not assert a difference when there is none[109]. However this tells us anything about the accuracy of the test where the null hypothesis is *not* rejected, which is achieved by lowering the Type II error rate. A low Type II error rate in a superiority trial means it unlikely that failure to reject the null hypothesis is a mistake, i.e. that some difference has not gone undetected. To speak loosely, a Type I error in a superiority trial is the mistake of accepting that there is a difference when there is none; a Type II error is the mistake of accepting no difference when there is one.

---

[108] Although it is often objected, notably by Bayesians, that the choice of Type I error rate at which rejection becomes justified is arbitrary, and that we should instead speak of hypotheses which are more or less probable (Howson and Urbach 1993).

[109] However, a low Type I or Type II error rate does not actually bear on the probability of a single trial, but only applies to the *long run*. That is, if the Type I error rate is 1%, this means that if we ran the same trial 100 times, we would expect 1 of them to give the wrong result. It says nothing about the probability of a particular trial being mistaken. I will ignore this problem with classical hypothesis testing for the purposes of my analysis.

In a non-inferiority trial, Type I and Type II errors, although defined in the same way, imply different things. In a non-inferiority trial, a Type I error would be the error of mistakenly rejecting the hypothesis of strict inferiority (negative difference) and accepting the hypothesis of equivalence or superiority. A Type II error in a non-inferiority trial, on the other hand, is the error of failing to reject inferiority when the experimental is either (roughly) equivalent or superior to the control treatment. To reduce the Type II error rate in a non-inferiority trial is therefore to make the trial very sensitive to differences: if the null is true and there is a difference (inferiority), then the trial will not reject it. Again speaking loosely, a Type I error in a non-inferiority trial is the mistake of accepting equality or superiority (when there is inferiority), while a Type II error is the mistake of accepting inferiority when there is equality or superiority.

The relationship between Type I and Type II error rates in superiority and non-inferiority trials is clearly explained with the help of a table:

13. Table 7.1: Classification of Possible Decisions Based on Hypothesis Tests (Adapted from Blackwelder (Blackwelder 1982))

		TRUE HYPOTHESIS			
		Null	*Alternative*	*Null*	*Alternative*
DECISION	*Reject*	**Type I error (α),** i.e. mistake of 'accepting' the alternative of *difference* (superiority)	correct	**Type I error (α),** i.e. mistake 'accepting' *no difference* (or superiority)	Correct
	Don't reject	correct	**Type II error (β),** i.e. mistake of 'accepting' the null alternative of *no difference* (or inferiority)	correct	**Type II error (β),** i.e. mistake of 'accepting' *difference* (inferiority)

With the conceptual issues out of the way, I will now examine whether, in fact, non-inferiority trials lack assay sensitivity.

7.4.2. Assay Sensitivity: A Property that Both Statistical Tests of Both Noninferiority and Superiority Trials Possess in Equal Measure

Recall that assay sensitivity, as we are viewing it in this section, is the ability of a trial to detect a difference between more and less effective treatments (rather than the ability to detect a difference between a placebo and a non-placebo). In this section, following Anderson, I will examine whether PCTs are inherently better at detecting any difference between more and less effective treatment.

It must be noted immediately that if ACTs are conducted as superiority trials, then they are clearly as competent at detecting a difference between more and less effective agents as PCTs. Hence, the assay sensitivity argument, on this interpretation, must be an argument against non-inferiority ACTs, which are not, and should not be, all ACTs.

Anderson (2006) postulates conditions required to assume assay sensitivity and claims that both PCTs and non-inferiority ACTs satisfy them equally. Using 'D' to denote "difference between intervention and control group", and T to denote "trial", the first three conditions are,

1. D
2. T indicates D
3. D → (T indicates D)

The first condition tells us that, ontologically speaking, there is a difference between the interventions being compared in the trial. The second tells us that the trial, in fact, indicated a difference. The third tells us that the trial would indicate a difference if there were one – that the trial 'tracks' the real difference. If this condition is not satisfied then a trial will indicate no difference when in fact there is one.

But the first three conditions tell us nothing about what the trial would indicate if there weren't a difference. That is, the three conditions alone are compatible with a trial indicating a difference when there is none. Thus, the conditions for concluding assay sensitivity needs to be augmented to include a fourth condition:

4. not-D → not (T indicates D)

"This condition tells us that the trial would not indicate a difference if there was not one" (Anderson 2006`, p. 76). In the real world, of course, the modality of the conditionals in (3) and (4) is not necessity – actual trials deal in probability[110]. If the fourth condition is not satisfied, then the trial will indicate that there is a difference when in fact there is none.

In a superiority trial, the four conditions for affirming assay sensitivity will be satisfied when:

1. There is a difference (the experimental treatment is superior to placebo).
2. The trial indicates a difference (there is a 'positive' result).
3. The Type II error rate is sufficiently low[111].

---

[110] Even then, a low error value does not mean that the probability of the error was low, but only that, if the trial were repeated many times, that the error would occur with a low probability. I will ignore this real problem with classical hypothesis testing for the purposes of this analysis since my argument in no way depends on it.

[111] Using the simplified terminology whereby 'D' signifies that there is a difference, and 'T' signifies that the trial indicates a difference, this can be shown quite easily. A Type II error in a superiority trial can be represented in symbolic logic as (D & not-T). If a Type II error has not been committed (or, more precisely, if we have reduced the probability of its occurrence) then we can assert its negation: not-(D & not-T), which is equivalent to (D → T), which is precisely

4. The Type I error rate is sufficiently low.

In a non-inferiority trial, the four conditions for affirming assay sensitivity will be satisfied when

1. There is a difference (the experimental treatment is superior to placebo).
2. The trial indicates a difference (there is a 'positive' result).
3. The Type I error rate is sufficiently low[112].
4. The Type II error rate is sufficiently low.

As is no doubt clear, as along as we reduce both the Type I and Type II error rates, *both* superiority and non-inferiority trials can be made equally assay sensitive by adjusting the Type I and Type II error rates. Anderson concludes, and he is surely correct: "Contrary to the assay sensitivity argument, there is not an absolute difference between PCTs and [non-inferiority] ACTs with respect to ... the assay sensitivity assumption" (Anderson 2006`, p. 78).

A proponent of this assay sensitivity could maintain that, even if non-inferiority trials are equally good at detecting differences, we still cannot infer from a 'positive' result of a non-inferiority trial to the conclusion that the experimental treatment is more effective than placebo. But this, of course, is the basic problem with ACTs (or, the 'first assay sensitivity argument') which was dealt with earlier.

Two final notes about this assay sensitivity argument. First, since it is about classical hypothesis tests of the different trials, it would not apply to statistical tests using Bayesian methods. The other arguments against ACTs, on the other hand, could apply no matter which type of statistical tests one employs. Similarly, this assay sensitivity argument applies exclusively (and not merely more strongly) to non-inferiority ACTs.

7.5. Conclusion: Assay Sensitivity Fails to Distinguish PCTs from ACTs

The basic problem with ACTs (the first 'assay sensitivity argument') only applies to cases where the effectiveness of the control treatments is in doubt. The 'sensitivity' argument only applies to cases where the control treatments have 'sensitivity' issues, which is, at best, a handful of treatments. The second 'assay sensitivity' argument is

---

the third condition for assay sensitivity. Similar results can obviously be obtained for showing the relationship between Type I errors in superiority trials and the fourth condition.

[112] This can be shown a similar way as it was for superiority trials.

limited in scope to non-inferiority ACTs, which are not always justified. Furthermore, the arguments against ACTs all apply either exclusively or to a greater degree, to non-inferiority ACTs. Non-inferiority ACTs, however, are not justified in as many cases as is often claimed. Perhaps more importantly, 'placebo' controls are more correctly viewed as treatments in and of their own right, it is clear that they suffer from all the potential problems of existing treatment controls. Since many 'placebo' controls are illegitimate, and since they sometimes perform erratically, the inference from a 'positive' result of a PCT to the claim that the experimental treatment was more effective than placebo also requires information external to the trial. In short, PCTs do not rule out the rival hypothesis that any observed effects were 'placebo' effects as unambiguously as must be assumed for the 'assay sensitivity' arguments to count against ACTs.

There is, of course, another potential reason to prefer PCTs, namely that they allegedly provide a measure of 'absolute effect size'. In the next chapter I will consider this other argument in more detail.

7.6. Chapter 6 Appendix A: Null and Alternative Hypotheses in Equivalence Trials

Clear discussions of hypothesis testing for equivalence and non-inferiority trials is not often found in major textbooks (Armitage, Berry, and Matthews pp. 636-7 is an exception) and where it is discussed in journals the discussion is often limited to non-inferiority trials, which are only one type of equivalence trials. Although the theoretical background is generally attributed to Dunnett and Gent (Dunnett and Gent 1977) and Blackwelder (Blackwelder 1982), this discussion follows Hwang and Morikawa (1999)[113] very closely because of its clarity of exposition and completeness.

There three ways to test for equivalence: using confidence intervals, using a single null hypothesis in a two-sided test, or using two null hypotheses in separate one-sided tests. I will limit my discussion to significance tests. For descriptions of the methods with confidence intervals see (Armitage, Berry, and Matthews 2002; Hwang and Morikawa 1999; Blackwelder 1982).

---

[113] Hwang and Morikawa developed their theory while participating in ICH E-10 expert working group.

Some formal vocabulary is required before describing the hypothesis tests in equivalence trials. Let T = test treatment, S = standard treatment control, P = 'placebo' control. The equivalence or non-inferiority margin, δ is the acceptable difference between T and S that is allowed in order to maintain equivalence. That is, T and S need not be exactly equal. Rather, they have to be similar within the acceptable margin δ – they must lie within the interval $[-\delta, \delta]$. The equivalence margin should generally be a fraction of the standard treatment control effect (over and above 'placebo'), Δ. This is self-evident. If δ approaches or surpasses Δ, then a result of 'equivalence' could mean that T is no more effective than 'placebo'. The margin must be chosen at the design stage and not *post hoc*. The ICH E-10 Guideline recommends that it be *at least* smaller than Δ (ICH 2000), but this may be too conservative (Hwang and Morikawa 1999`, p. 1207). Hwang and Morikawa recommend that the margin be less than ½ Δ in order to be useful. At the same time it cannot be too small or the sample size will be impractically large.

For example, consider a trial of a new drug and a standard drug for hypertension. The primary endpoint will be supine diastolic blood pressure (SDBP, measured in mmHg). The minimum drug effect size Δ (standard drug – 'placebo') in mean reduction from baseline in SDBP in many 'placebo' controlled trials was approximately 10 mmHg. A choice of $\delta = 3$ mmHg was considered reasonable based on clinical relevance and statistical judgment – it assured that some clinically acceptable drug effect would be maintained without having an unrealistically large sample size.

In an equivalence trial, the null hypothesis (which investigators will try to reject) is that the difference is neither superior nor inferior to the upper and lower limits of the limits of the equivalence margin. The alternative hypothesis is that the difference lies within the range of acceptable difference. That is:

(i) $h_0: T - S \leq -\delta$ or $T - S \geq \delta$

vs. $h_a: -\delta < T - S < \delta$

Alternatively, equivalence can be tested with a pair of one-sided null hypotheses:

(ii) $h_0: T - S \leq -\delta$ vs. $h_a: T - S > -\delta$ and

$h_0: T - S \geq \delta$ vs. $h_a: T - S < \delta$

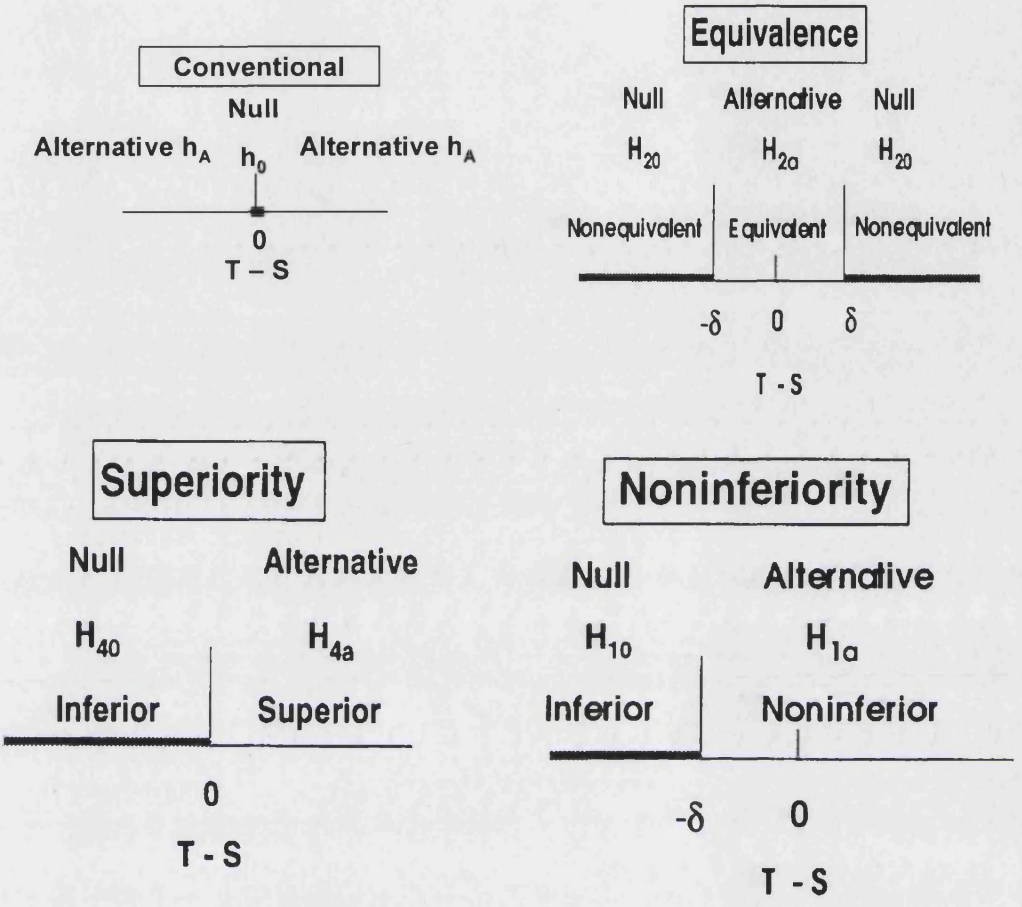
Non-inferiority trials can either use method (i) in a one-sided test, or the first pair of hypotheses in (ii).

Note that compared to 'placebo' controlled *superiority* trials, the null and alternative hypotheses in non-inferiority trials are essentially reversed. In a 'placebo' controlled trial, the null and alternative hypotheses are:

(iii) h_0: $T - S \leq 0$ vs. h_a: $T - S > 0$

The relationship between the null and alternative hypotheses in 'conventional', equivalence, superiority and non-inferiority trials is best explained pictorially (see figure below).

14. Figure 7.1: Null and Alternative Hypotheses for Equivalence, Noninferiority, and Superiority Trials (adapted from Hwang and Morikawa, pp. 1210-11).



7.7. Chapter 6 Appendix B. The Ethical and Methodological Problems with Placebo Controlled Trials

As I already argued in detail in chapter 4, the methodological problem with PCTs is that it is difficult, and sometimes impossible to design 'legitimate'placebo' controls that actually control for all 'non-characteristic' features of the experimental treatment. 'Placebo' controlled trials also fail to deliver practically useful knowledge. What the average patient or dispensing physician wants to know is which intervention, from among the available alternatives, is most effective (or has fewest side-effects, or is easiest to take, or is cheapest). The average patient or dispensing physician standardly has little interest in knowing whether some new treatment is simply better than 'placebo', and wishes to know how effective the experimental treatment is relative to the best existing treatment. It is assumed, of course, that at least some existing

treatments are more effective than 'placebo'. This practically useful knowledge may be provided directly with ACTs.

It is beyond the scope of this work to discuss the ethical debate over the use of 'placebo' controls in any detail. What is important for the purposes of this work is to note that the justification for using 'placebo' controls even where there is an established treatment is that PCTs are methodologically superior to ACTs[114]. The tension between ethics and methodology here is apparent if we examine the change in the World Medical Association's *Declaration of Helsinki.* The Declaration initially (1964) condemned the use of PCTs where known standard intervention exists:

> The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current ... methods. This does not exclude the use of placebo, or no treatment, in studies where no proven ... method exists" (WMA 1964)

Yet it is clear from the wording of the revised 2001 Declaration that methodological considerations motivated the revision:

> a placebo controlled trial may be ethically acceptable even if proven therapy is available, under the following circumstances:
>
> - Where for compelling and scientifically sound methodological reasons *its use is necessary to determine the efficacy or safety of a prophylactic, diagnostic, or therapeutic method* ... (WMA 2001)

The Declaration, however, is silent when it comes to specifying the "compelling and scientifically sound methodological reasons". A list of these reasons can, however, can be found in the International Conference on Harmonization E10 (ICH E10) document, produced and endorsed by the regulatory bodies of the United States, the European Union, and Japan. According to this document, there are three relative disadvantages of ACTs. First, ACTs do not provide a direct measure of effect size whereas PCTs do. Second, ACTs do not always possess 'assay sensitivity', whereas PCTs do. Third, ACTs require a larger sample size than PCTs.

---

[114]Even the superficially uncontroversial argument that placebo controls are unethical when well-established treatment is available has been challenged on the grounds that the ethical duty of the dispensing physician to provide the best possible treatment is different from the ethical duty of the investigator (see for example Miller and Brody: 2002). Not everyone was convinced by this argument and it has been the subject of controversy (Mann 2002; Ackerman 2002; Weijer and Glass 2002; Weijer and Miller 2004).

8. Chapter Eight. The Assumption of Additivity in Placebo controlled trials: Exploring the Myth that Placebo controlled trials Provide a Measure of Absolute Effect Size

The power attributed to morphine is then presumably a placebo effect plus its drug effect
- (Beecher 1955)

This principle [of the Composition of Causes] ... by no means prevails in all departments of nature
- (Mill 1843[1973])

8.1. Introduction

In this chapter I will evaluate whether 'placebo' controlled trials are superior to 'active' controlled trials on the grounds that only the former provide a measure of 'absolute effect size'.

The International Conference on Harmonization guidelines, produced by the regulatory authorities of the United States, Japan, and the European Union states that an advantage of placebo controlled trials is that they measure 'absolute' efficacy and safety (ICH 2000`, p. 18). 'Active' controlled trials, on the other hand, allegedly do not have this property:

> Even when assay sensitivity is supported and the study is suitable for detecting efficacy, there is no direct assessment of effect size, and there is also greater difficulty in quantitating safety outcomes (ICH 2000`, p. 18).

The absolute effect can allegedly be estimated by taking the "difference in outcome between the active treatment and placebo groups within the trial" (ICH 2000`, p.13). What is meant, of course, and this is worthwhile pointing out from the outset, is that placebo controlled trials do not provide an absolute measure of the effectiveness of the experimental treatment, but rather (allegedly) of the effects of the *characteristic*[115] features of the experimental treatment. More precisely the outcome in the experimental treatment group, T, is due to:

(1) 'Natural history' or 'spontaneous remission' (*n*). I will call the effects of natural history will then be N.

---

[115] The terms 'characteristic' is borrowed from Grünbaum (1986) and discussed at length in chapter 3. For present purposes, it is sufficient to note that the 'characteristic' features are the 'non-placebo' features, such as the chemical fluoxetine in treatment involving Prozac.

(2) Beliefs and expectations (*b*). I will call the effects of beliefs and expectations B.

(3) Characteristic features of the treatment (*c*). Call the effects of these features C

Many ailments, including the flu, mild depression, pain, sprained limbs, stomach poisoning, often go away in time without any outside intervention at all. This obvious fact has often been ignored by scholars of the 'placebo' effect, who take the outcome after administering a 'placebo' pill to be the effects of beliefs and expectations. The outcome in the 'placebo' control group, even if apparently dramatic, might have nothing whatsoever to do with the effects of expectations and beliefs and could merely be the effect of 'natural history' or 'spontaneous remission'. People can recover quite dramatically without any treatment at all. Next, there are the potential effects of beliefs and expectations that one is being treated with a powerful, 'non-placebo' treatment. For simplicity, I will represent the effects of these two 'non-characteristic' features (N and B) as one variable I. Finally, there are the characteristic features of the experimental treatment, i.e. fluoxetine in treatment with Prozac.

The outcome in the 'placebo' control group (P), on the other hand, is (or at any rate should be) due only to the 'non-characteristic' features, which include both natural history (*n*), and beliefs and expectations (*b*).

The effect of the characteristic features of the treatment (*c*), i.e. fluoxetine in treatment with Prozac, is calculated by subtracting the average outcome in the placebo group from the average outcome in the experimental group. This difference allegedly provides an absolute measure of the characteristic features of the experimental treatment (in the example, fluoxetine). That is, $T - P = C$.

But in order for this equation to hold, we must assume additivity. The assumption of additivity amounts to the claim that the outcome in the experimental group, $T = I + C$, while the outcome in the 'placebo' control group, $P = I$. Quite obviously, if this assumption holds then $T - P = (I + C) - (I) = (C)$, and the claim that placebo controlled trials provide an 'absolute measure of effect size' can be justified.

In this chapter I will argue that the assumption of additivity is a substantive one that must be justified rather than simply assumed. Before we accept the claim that placebo controlled trials provide an absolute measure of characteristic effectiveness, the rival hypothesis of interactions must be taken seriously. If the characteristic and non-characteristic features interact, then it is incorrect to represent the outcome in the experimental group as $T = I + C$, and there will not be a straightforward sense in which placebo controlled trials provide an absolute measure of the characteristic effects.

Although interactions between drugs have been taken seriously for several decades now, the possibility of interactions between characteristic and non-characteristic treatment features has all but been ignored.

To anticipate, I begin with a description of Mill's Principle of the Composition of (mechanical) causes to describe the assumption of additivity in more detail. I then outline *prima facie* reasons to question the assumption in placebo controlled trials, and provide evidence of some cases where additivity does not seem to hold. Next, I note further reasons to dispute the claim that placebo controlled trials provide a measure of 'absolute' effect size. Even if additivity can be justified (which it no doubt sometimes can), placebo controls still merely provide a measure of effectiveness that is relative to a particular trial. A consequence of non-additivity in placebo controlled trials is that their external validity can be questioned. I conclude that 'placebo' controlled trials do not provide a measure of absolute effect size unless the rival hypothesis of interactions has been ruled out. The possibility of interactions between characteristic and non-characteristic features of a treatment has been largely ignored; when considered, there are good reasons to assert that interactions are, in at least many cases, much more than a possibility.

8.2. From Mill's Principle to the Implicit Assumption of Additivity in Placebo Controlled Trials

The assumption of additivity, although no doubt applied long before Mill, was eloquently expressed by Mill's Principle of the (mechanical) Composition of Causes[116]. Here, one cause, such as a vertical force acting on a billiard ball, adds (of course in this case 'adds' in the vector sense) to another cause, such as a horizontal force acting on a billiard ball, to produce the resultant force.

---

[116] Although use of the term 'cause' could be construed as introducing unnecessary ambiguity, Gillies has recently (2005) argued that the notion of cause plays a legitimate and central role in medicine. More specifically, Gillies argues that although Russell was correct that the term 'cause' has no place in advanced physical sciences, that he was wrong when he claimed that the term had no place in *all* advanced sciences, in particular medicine. The notion of causes, Gillies argues, plays a central role in the diagnosis, prevention, and treatment of disease. Gillies moves on to defend an 'action-related theory of causality' that is similar, but relevantly different to Menzies and Price's theory (Gillies 2005).

Mill explains the principle of the Composition of Causes with reference to the principle of Composition of Forces.

> If a body is propelled in two directions, by two forces, one tending to drive it to the north and the other to the east, it is caused to move in a given time exactly as far in both directions as the two forces would separately have carried it; and is left precisely where it would have arrived if it had been acted upon first by one of the two force and afterwards by the other. This law of nature is called, in dynamics, the principle of the Composition of Forces: and in imitation of that well-chosen expression, I shall give the name of the Composition of Causes to the principle which is exemplified in all cases in which the joint effect of several causes is identical with the sum of their separate effects (Mill 1973[1843] I.v.1).

In short, assuming the principle of Composition of Causes allows us to deduce the overall effect by performing (vector) addition on the component causes. "Now, if we happen to know what would be the effect of each cause when acting separately from the other, we are often able to arrive deductively, or *a priori*, at a correct prediction of what will arise from their conjunct agency" (Mill 1973[1843] I.v.1). Since the Principle of the Composition of Causes is essentially the same as the assumption of additivity that is often unthinkingly made regarding placebo controlled trials. I will use both terms interchangeably.

Mill was careful to note that his Principle is not universal. He was explicit that although it applied to the composition of 'mechanical' forces, it generally did *not* for chemical or biological causes.

> The chemical combination of two substances produces, as is well known, a third substance with properties different from those of either of the two substances separately, or of both of them taken together. Not a trace of the properties of hydrogen or oxygen is observable in those of their compound water (Mill 1973[1843] I.v.1).

That is, the Principle did not generally hold in chemistry. In biology Mill claimed that there is even less reason for the Principle to hold.

> If this be true of chemical combinations, it is still more true of those far more complex combinations of elements which constitute organized bodies; and in which those extraordinary new uniformities arise, which are called the laws of life (Mill 1973[1843] I.v.1).

Mill illustrates with the example of the tongue.

> The tongue, for instance, is, like all other parts of the animal frame, composed of gelatine, fibrin, and other products of the chemistry of digestion, but from no knowledge of the properties of those substances could we ever predict that it could taste, unless gelatine or fibrin could themselves taste; for no elementary fact can be in the conclusion, which was not in the premises (Mill 1973[1843] I.v.1).

In short, biological and chemical causes seem to *interact* more often than they add. Given that medical treatments, or at any rate pharmacological treatments, are biochemical rather than mechanical, it seems that we have good *prima facie* reasons to question whether additivity holds between the characteristic and non-characteristic features of a treatment. Given Mill's intuitively uncontroversial arguments that his Principle only holds (in general) for mechanical causes, and given that on the face of it non-characteristic and characteristic features do not appear to be mechanical, we might think that interaction rather than additivity should be the default position, or at least as a possible option when it comes to medicine.

Cartwright, however, appears to take the view that in the absence of reasons to believe otherwise, that Mill's Principle holds:

> [O]ne looks for independent evidence that an interaction is occurring, and some account of why it should occur between these variables and not others, or at these levels and not others. The chemistry examples are a good case. One does not just say the acid and base interact because they behave differently together from the way they behave separately; rather, we understand already a good deal about how the separate capacities work and why they should interfere with each other in just the way they do (Cartwright 1989`, p. 165)

Cartwright seems to place the burden of proof on those who wish to reject additivity. However, it is quite clear that though additivity may hold in general, interactions are commonplace in (non-mechanical) physics, biology, and chemistry. For instance, threshold interactions are common in biology. Someone's height may depend on their genetic constitution, but without a threshold level of nutrition, whatever genetic tendency they have may not manifest itself. Schizophrenia is often considered to be a threshold trait. Many factors are thought to contribute to the development of schizophrenia but only when these factors reach a threshold level do they produce the condition. Individually, the factors do not have any effects on schizophrenia: "there exists a threshold such that individuals above the threshold have schizophrenia and those below do not" (Sullivan, Kendler, and Neale 2003`, p. 1188).

Even in medicine, at least where there is more than one characteristic feature, investigators take great pains to determine whether there are interactions. Whenever there are more than two drugs involved in a trial, statistical tests for interactions between the drugs are almost always made. The test involves calculating whether changing one variable produces a differential change in the other (Armitage, Berry, and Matthews 2002`, p. 244). For example, the ISIS-2 study showed that aspirin and

streptokinase each improve outcomes in myocardial infarction, and that the combination of both these drugs has an additive therapeutic effect (Baigent et al. 1998).

To cite another example, hypericum perforatum (St. John's Wort), a herb that has been shown in some trials to be as effective as pharmaceutical antidepressants, seems to interact with many other drugs. St. John's Wort activates a nuclear receptor called the pregnane X receptor (PXR). PXR is a ligand-activated transcription factor that induces a number of xenobiotic-metabolizing enzymes and transporters including cytochrome P4503A4 (CYP3A4) in humans. Because CYP3A4 alone metabolizes about 60% of all clinically relevant drugs, induction of CYP3A4 may result in the rapid elimination of these drugs and a consequent reduction in drug efficacy (Choudhuri and Valerio 2005). Several studies have shown that St. John's Wort in fact metabolizes the characteristic features of certain contraceptives (Hall et al. 2003; Pfrunder et al. 2003)[117]. In short, the possibility of interactions between two 'active' substances is considered very seriously.

Interaction between two drugs, however, does not impinge on the assumption of additivity between the characteristic and non-characteristic features of a treatment. Nonetheless, I wish to argue that additivity between characteristic and non-characteristic factors (henceforth simply 'additivity') is a substantive assumption – indeed a very strong one. Theoretically, the overall effect could be *any* combination of the various component causes, including $T = (I + C)^2$, or $T = \exp(I)/C$, or indeed anything else. If there are interactions between non-characteristic and characteristic features, then it will be impossible to 'tease out' the effect of the characteristic treatment features by subtracting the average outcome in the 'placebo' control group (P) from the average outcome in the experimental treatment group (T).

Of course, even if there are interactions, a placebo controlled trial will provide a measure of how much the characteristic features increased the overall effectiveness. That is, if we define characteristic effectiveness more generally as the difference between the average outcome in the experimental group less the average difference in the placebo control group, we will get an estimate of whether and by how much the characteristic features 'add value'. But this measure is surely not *absolute*, since if there are interactions, then the measure will be strictly relative to the particular trial. For

---

[117] A more recent study, however, suggests that the effectiveness, measured by contraceptive efficacy, is not affected by hypericum (Fogle et al. 2006).

instance, in the imaginary case where $e(n, b, c) = e(n) + e(b) + De(c)$, where $D = 1$ for $e(b) < e(c)$, and $D = 0$ otherwise, changing the strength of the beliefs/expectations can completely change the measure of characteristic effectiveness, $e(c)$.

To sum up this section, the implied assumption of additivity in placebo controlled trials is that the characteristic and non-characteristic 'causes' combine according to Mill's Principle of the Composition of Causes to produce the overall effect. Since, however, there are good reasons to doubt whether additivity holds in biology and chemistry, there seems to be good reason to demand that additivity in placebo controlled trials be justified rather than simply assumed. In short, I have presented a rival hypothesis to the claim that placebo controlled trials provide a measure of absolute effect size.

My next step will then be to argue that the rival hypothesis of interactions, in addition to being possible, is also plausible. In the next section, I will consider whether there is any empirical evidence that bears on this issue.

8.3. Evidence for Interactions

The most striking thing about evidence for interactions between characteristic and non-characteristic features is that it is sparse. The lack of evidence might be due to the fact that besides two interesting papers (Kleijnen et al. 1994; Kirsch 2000), little research has been devoted entirely to questioning the assumption. The other reason why there is little evidence for interactions between expectations and characteristic features is that it is difficult to gather. In order to test whether there are interactions between non-characteristic and the characteristic features of a treatment process, we need examples where we had more than one 'dose' or 'strength' of the non-characteristic features in the study (I will explain how this can be done shortly). If changing the 'dose' of the non-characteristic features produces a differential change in the characteristic features, then we can conclude that there are interactions. For example, if the characteristic effectiveness were very high with a low 'dose' but very low with a high 'dose' of, say, expectations, then we would have evidence for interactions. With this in mind, the following examples are not intended to show that additivity never holds, but only to suggest that additivity cannot simply be assumed.

Studies that have used the balanced placebo design, described in more detail in chapter 5, provide examples where there are both high and low 'strengths' of

expectations. Assuming that the other non-characteristic features, namely natural history and regression to the mean, remain the same, then changing the strength of expectations changes the strength of the non-characteristic features. Briefly, the balanced placebo design uses four groups. Two groups are told they are receiving the experimental treatment (say nicotine gum), but only one actually receives it (the other receives 'placebo'). The expectations in these groups can be said to be high since they expect the experimental treatment rather than 'placebo'. Then, the other two groups are told they are receiving 'placebo', but one of the two actually receives the 'placebo' (the other receives the experimental treatment deceptively). The expectations in these latter groups will be lower since they expect to receive the 'placebo'.

The balanced placebo design allows us to test for interactions in the following way. The characteristic effectiveness can be calculated by subtracting the effect of being given the placebo from the effect of being given the experimental treatment for two 'strengths' of expectations. If the characteristic effectiveness changes depending on whether the 'strength' of the expectations is high or low, then we can conclude that there are interactions.

Hughes *et al.* (1989), found that although the characteristic effects of nicotine gum were significant at 'doses' of expectations, that there were no such effects at higher 'doses'. The trial involved 77 smokers[118] who quit smoking and were assigned one of the following 6 groups who were each told something different:

(1) told given nicotine gum[119], and given nicotine gum

(2) told given nicotine gum, and given 'placebo' nicotine gum

(3) told given 'placebo' nicotine gum, given nicotine gum

(4) told given 'placebo' nicotine gum, given 'placebo' nicotine gum

(5) not told whether given nicotine or 'placebo' gum, given nicotine gum

(6) not told whether given nicotine or 'placebo' gum, given 'placebo' nicotine gum

This trial design allows us to estimate the characteristic effects of nicotine gum under different 'strengths' of expectations (and hence, we assume, non-characteristic features).

---

[118] In order to be eligible for the trial, participants had to have smoked 10 cigarettes of at least 0.5mg nicotine per day for at least 1 year, fulfil certain criteria for tobacco dependence, believe that nicotine gum could relieve withdrawal symptoms, and not be taking concomitant medications (Hughes et al. 1989).

[119] 'Nicorette', 2mg nicotine, Merrell-Dow Pharmaceuticals.

It is safe to assume that being told one is receiving the experimental treatment increases the 'strength' of beliefs/expectations. These conditions were emulated in groups (1) and (2), which are the conditions of routine practice, where treatments are usually administered as having established 'non-placebo' effectiveness. Meanwhile, the conditions in groups (5) and (6) emulated the conditions of a typical double masked trial. The strength of expectations in trial conditions will typically be lower than in the case where the participants are (albeit deceptively) informed that they are being given the experimental treatment rather than the 'placebo'. Finally, the expectations will be least strong if one is told outright that one is being given the placebo. These conditions are emulated in groups (3) and (4).

The participants provided their informed consent to have a 50/50 chance of receiving nicotine or placebo gum, and that they might or might not be told the contents of their gum. They were *not* told that they could be deceived (Hughes et al. 1989). They were all told to use the gum when the urge to smoke occurred. To ensure that participants who were told they received placebo (groups (3) and (4)) sampled the gum and tried to quit, they were repeatedly told that studies indicate that 'placebos' sometimes help people quit.

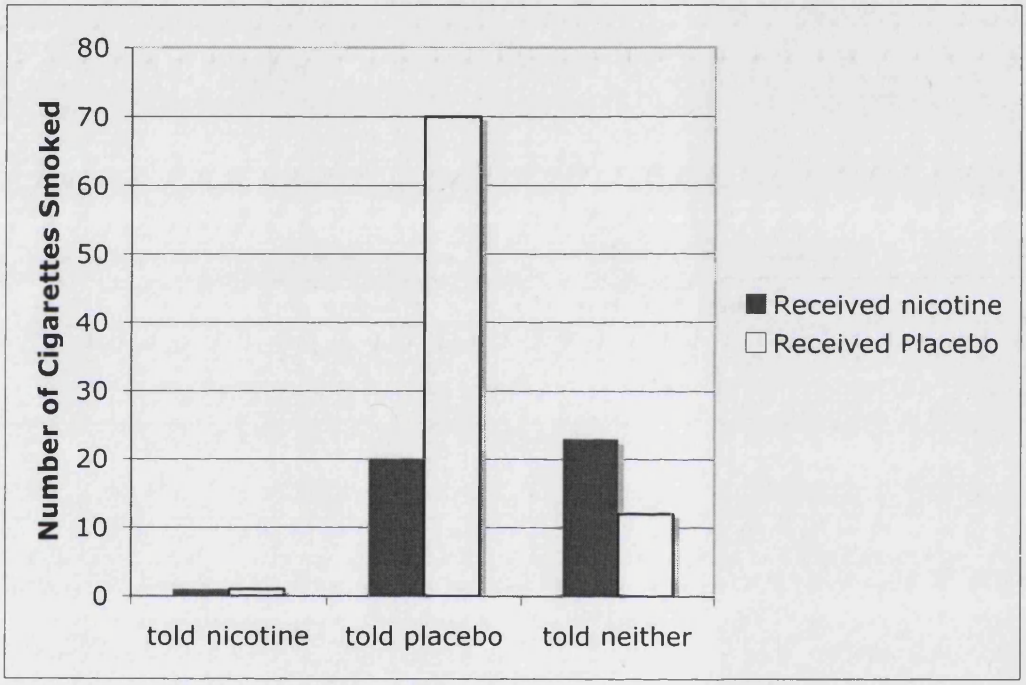
The outcome measures were proportion of participants who smoked no cigarettes during the week, proportion who smoked on fewer than 2 days per week, number of days smoked per week, and number of cigarettes smoked per week. These were calculated based on assessments measured 1 and 2 weeks after they attempted to quit, and were measured in three ways. First, the participants self-reported how many cigarettes they smoked. Second, a designated observer (usually a spouse) reported the participant's smoking habits during the week. Finally, a breath sample of carbon monoxide was taken to verify claims of complete abstinence.

At the end of the trial, participants were also asked whether they thought they had been deceived; 29% thought they had never been deceived, 63% considered deception but were unsure, while 8% believed they were deceived. The distribution of these figures was the same across groups. In short, the deception appeared successful.

Unsurprisingly, the overall effect was strongest in the cases where the participants were told that they received real nicotine gum ($P = 0.01 - 0.08$). The evidence for non-additivity, however, is provided by the fact that, in the groups that were told they received nicotine gum, the difference between those who received 'placebo' gum and those who received nicotine gum was much smaller than the

difference between those who actually received 'placebo' or nicotine gum in the other groups. This was repeated for all the outcome measures. For at least two of the measures (number of days smoked, number of cigarettes smoked), it seems that the effects of the beliefs/expectations allowed hardly any room at all for drug-induced improvement. The statistical evidence for interactions between instructions (what they were told, which impacts on expectations) and the overall outcome was significant (P = 0.01). This is most easily understood with the help of a chart:

15. Chart 8.1: Smoking behaviour by instruction and drug group. The results are cumulative across the two weeks where assessments were made. Reproduced from Hughes (1989).



From the figure it is clear that as the 'strength' of expectations increase, there is less and less room for drug induced improvement. In conditions similar to those of routine practice, there is no room for drug induced improvement at all, on at least this outcome measure. The results followed the same trend for the other outcomes.

In another example, Bergmann and colleagues in France conducted an interesting study immediately before informed consent became a requirement. Fourty-nine patients with mild to moderate cancer pain were randomly selected to receive or not receive information about participating in a randomized, double masked crossover trial of naproxen and 'placebo'. In a randomized crossover design, a random process determines which of the experimental treatment (i.e. naproxen) or control (i.e. 'placebo') comes first. In this case a random process determined whether a patient

received naproxen on the first morning, and 'placebo' on the next, or vice versa. The treatments were delivered in double masked conditions.

Half the patients (25) were randomized to participate in the trial without any information, i.e. without providing their informed consent. At the time, French regulations did not require systematic informed consent, and it was approved by the French ethical committee, but such a trial would never be allowed in most countries including France today. The uninformed patients who were given naproxen were told they would be given naproxen, while the uninformed patients who were given placebo were treated 'covertly'.

> Twenty-five patients therefore received in a randomised order 500 mg per os of naproxen or matched placebo on the first morning and the crossover drug on the second morning with information about naproxen but no prior information about the crossover trial and the placebo (Bergmann, Chassany, and Gandiol 1994, p.42)

The other half (18 – six had refused to participate when presented with the choice to consent) provided their consent to receive either placebo or naproxen under double masked conditions. Informed consent consisted of "detailed information relating to the crossover placebo-controlled study" (Bergmann, Chassany, and Gandiol 1994, p.42).

The baseline characteristics (age, sex, type of cancer, location of pain, and pain level before study), were similar (Bergmann, Chassany, and Gandiol 1994, p.44). The expectations in the uninformed group can fairly said to be lower than the expectations in the informed group, since the latter believed they had some chance of being treated with the 'real' painkiller.

The outcome was pain reduction after 30, 60, 120, and 180 minutes after treatment, measured on the visual analogue scale[120]. The reduction in pain was higher in both informed groups ($P = 0.001$)[121]. Interestingly, the *difference* between experimental

---

[120] The visual analogue scale is a pain measurement tool consisting of a straight line, usually 10cm in length. One end of the line represents 'no pain', while the other end represents 'very severe pain'. Patients are asked to place a mark to indicate their level of pain.

[121] More specifically, naproxen was more effective than placebo in both groups ($p = 0.001$). For naproxen and placebo, the characteristic analgesic effect was better in the informed group compared to the uninformed group ($p = 0.012$). The difference in therapeutic activity between naproxen and placebo was moderately higher in the uninformed patients ($p = 0.08$) (Bergmann, Chassany, and Gandiol 1994).

and 'placebo' groups was different in the informed and uninformed groups – it almost twice as great in the uninformed group (see table below).

16. Table 8.1: Mean reduction in pain (in millimetres on visual analogue scale) for naproxen and placebo under informed and uninformed conditions

	Informed Consent (n=18)			*No information (n=25)*		
Time (min)	naproxen	placebo	**difference**	naproxen	placebo	**difference**
30	8.4	2.2	**6.6**	3.6	-1.8	**5.4**
60	14.9	10	**4.9**	5.9	-3.4	**9.3**
120	21.9	15.7	**6.2**	10.8	-5.9	**16.7**
180	22.1	19.2	**2.9**	5.3	-8.5	**13.8**

Naproxen is known to have its maximal effect after 2-4 hours, which explains the relative lack of difference after 30 and 60 minutes. To be sure, the sample sizes were quite small; the variation, which I omitted from the table for simplicity, was quite high, and the statistical test for interaction (non-additivity) did not reach statistical significance ($P = 0.08$). However, interactions may have become apparent if the sample size were larger. These results suggest that the potential for characteristic effects tapers off as the non-characteristic effects increase.

Both examples which show apparent evidence for interactions can be discussed further and indeed questioned. The follow up time for the nicotine gum study, for instance, was quite short. However, the aim of this section was not to provide conclusive evidence for interactions, but merely to indicate that additivity may not always hold.

In fact, the results of the two examples above are hardly surprising. Many ailments can only be relieved by so much: there is a maximum possible effect. If the expectations alone produce this maximum effect (or something close to it), then there will be little room for characteristic drug induced improvement. If, for example, someone is dehydrated, then drinking enough water to become hydrated will produce the maximum effect; if the person drinks more water they will not become 'super' hydrated. The mechanisms that account for a maximum drug response are often understood (Aronson 2007). To oversimplify: the maximum response is related to the maximum number of receptors cells have for the drug to attach-to. Once the receptors have all been occupied, there is no further room for improvement. In short, the fact that the dose-response curve 'flatlines' beyond a certain dose is no mystery.

If we combine the fact that dose-response curves 'flatline' with the fact that at least some ailments – such as pain, depression – are responsive to expectations and beliefs, then it is possible that the characteristic effects taper off as the strength of the expectations features increase. This is because the increased expectation effects leave little room for characteristic feature-induced improvement.

It is possible, of course, that expectations and characteristic features interact in other ways. The examples discussed so far were cases where the non-characteristic features 'antagonized', or reduced the effect of the characteristic feature. It is possible, of course, for the non-characteristic and characteristic features to combine synergistically. One type of synergistic interaction is a threshold interaction, whereby non-characteristic features could 'potentiate' the characteristic features or vice-versa. For example, it is commonly claimed that Freud insisted on a hefty fee not because he was greedy but because he thought that the fee would act as a catalyst for what are commonly thought of as the characteristic features of Freudian psychoanalysis (Grünbaum 1986`, p. 24).

Synergistic interactions, like 'flatlining', can also be explained, at least in principle. There is evidence that placebos for pain increase the levels of endogenous opioids (Benedetti and Amanzio 1997; ter Riet et al. 1998). The increased opioid level could stimulate interaction with the characteristic features to increase the effects by interacting synergistically with the active treatment.

Interactions could also have a more dramatically 'antagonistic' effect than simply 'flatlining' (where increased expectations seemed to reduce the characteristic effects). Levine and Gordon (1994) found that the strength of the expectation effects could antagonize (have a negative effect on) the characteristic effects of the drug, changing the results of the study from an apparently 'positive' to a 'negative' finding of characteristic effectiveness.

In the study ninety-six patients who had undergone surgery for impacted molars were delivered naloxone (an opiate antagonist which will reduce the effect of the operative anaesthesia, and *increase* pain) or naloxone 'placebo' (a saline solution, which the authors refer to as 'vehicle'), or morphine. All treatments were delivered via an indwelling intravenous line. They were administered naloxone or 'vehicle' either 'openly', i.e. by a physician at the bedside, or 'covertly', i.e. by a physician in another room, or by a pre-programmed machine. In short, the treatments were delivered in three ways:

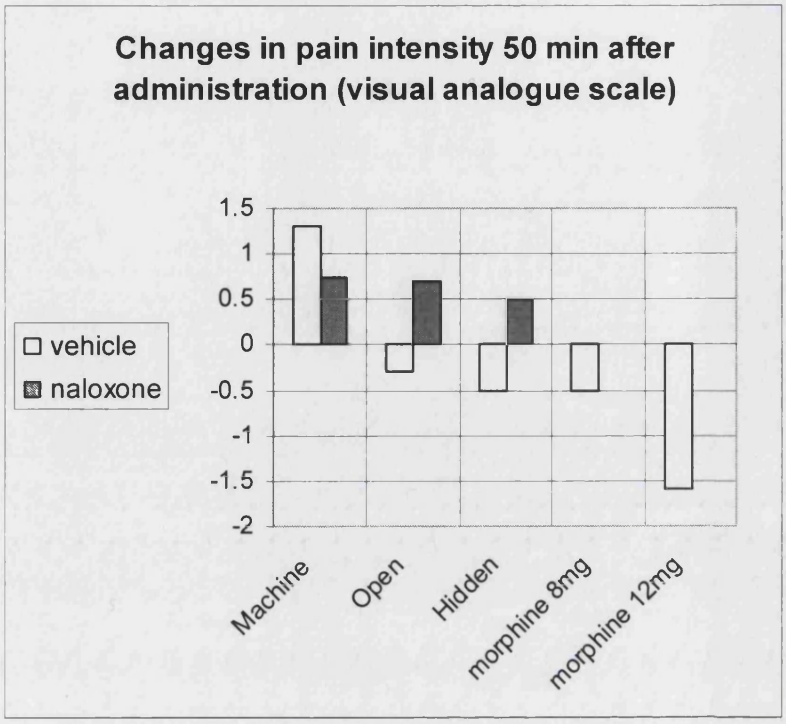
(1) Naloxone or vehicle from a physician at the bedside ('open' or 'overt' administration), in double masked conditions (i.e. neither physician nor patient was aware whether the substance was naloxone or 'placebo')
(2) Naloxone or vehicle from a physician in a different room ('hidden' or 'covert' administration), also in double masked conditions.
(3) Naloxone or vehicle from a pre-programmed infusion pump ('machine' infusion), also under masked conditions.
(4) Either 8mg or 12 mg of morphine.

The morphine was administered in order to measure the relative potency of the 'placebo'. The substances were all delivered three hours after onset of anaesthesia. The outcome was pain reduction after 50 minutes, and was measured using the 10cm visual analogue scale described above. The authors do not state whether the assessment was conducted in masked conditions.

Many interesting hypotheses were tested in this study. For example, the hypothesis that hidden administration of naloxone might be more effective than machine administration, which would mean that subtle cues can transmit expectations and beliefs (it turns out that hidden administration was more effective than machine administration). What is relevant for present purposes, however, is the relationship between the different levels of expectations (which are enhanced in the 'open' group) and naloxone itself.

The results indicated significant interactions ($P < 0.01$), and that "open infusion of naloxone produced an increase in pain whereas open infusion of vehicle produced a decrease, suggesting that there is a naloxone-antagonizable component of placebo-induced analgesia" (Levine and Gordon 1984, p. 755). The results are best explained with the help of the chart (see chart).

17. Chart 8.2: Changes in pain intensity (reproduced from Levine and Gordon, 1984)



Examination of the chart makes it quite clear that the expectation/belief effects seem to have an adverse effect on naxolone rather than an enhancing effect. This suggests that there is a non-additive relationship. If additivity held, then we would expect the characteristic effects (calculated from subtracting the outcome in the 'placebo' group from the effects in the experimental treatment group) of naloxone to be the same regardless of the expectations. In fact, the characteristic effects of naloxone with machine infusion (low expectations) appear to be *negative*, while the characteristic effects of naloxone with open infusion appear positive.

One might be tempted to criticise the naloxone study in the following way. Post-operative patients still under the after-effects of anaesthesia might not understand what naloxone is supposed to do. Rather, they might assume that any treatment given to them by a physician will help them and decrease the pain. Certainly the concept of an 'opioid antagonist' is unfamiliar to the average patient, and many patients might simply assume that their physicians are trying to do 'good' (i.e. reduce pain) rather than 'harm' (i.e increase pain). If patients mistook potential treatment with naloxone to be treatment with something that reduced pain, then they would expect their pain to decrease, rather than increase, and this would seem to explain the results. The authors do not consider this possibility.

It might also be objected that even if additivity between expectations and characteristic features does not hold in the examples above, it does hold in many other cases. Furthermore, the examples I used relied on real expectation effects. If, as was suggested in chapter 5, some ailments are not 'placebo' responsive (i.e. not responsive to placebo effects), then it would seem that the very possibility for interactions is precluded.

My response here is similar to my response to earlier objections to the alleged evidence for interaction. That is, the modest intent of this section was to point out that additivity between expectations and characteristic features might not always hold and therefore that additivity could not simply be assumed. If there are good reasons to suppose that there are no interactions (i.e. in cases where it is unlikely that there are expectation effects at all), then this would count as some justification for assuming additivity.

Other studies that used the balanced placebo design and that provide evidence for interactions include (Hughes et al. 1989; Rohsenow and Marlatt 1981; Ross and Pihl 1989; Mitchell, Laurent, and de Wit 1996), while other articles discuss the possibility of interactions without conducting any primary research (Kaptchuk 2001; Kleijnen et al. 1994)

Before moving to the next section, where I will argue that even if additivity always held, placebo controlled trials still do not provide a measure of absolute effect size, I will comment on how interactions bears on the external validity of trials.

8.3.1. Interactions reduce the external validity of placebo controlled trials

In order to generalize the results of a clinical trial, we must assume that the conditions of the trial (such as patient characteristics) are relevantly similar to the conditions in routine practice. Before generalizability can be assumed, the rival hypotheses that the patient characteristics are relevantly different, and the trial conditions were relevantly different, must be ruled out. The possibility of interactions presents an additional reason to worry about external validity. That is, the different level of beliefs and expectations in routine practice could affect the characteristic effect of the drug quite independently of possible different patient characteristics and other trial conditions.

It is safe to assume that the level of expectations in routine practice are higher than in a double masked, placebo controlled trial. In routine practice treatments are

delivered by physicians who are (hopefully) confident that the treatment is helpful and that it is not harmful. In a double masked placebo controlled trial, the treatments are delivered with some measure of doubt. Neither the physician nor the patient will know whether a particular treatment is the experimental intervention (which, because it is experimental, has presumably not established its effectiveness as securely as interventions given in routine practice)[122], and which is the 'placebo'. If there is a 'flatlining' effect, then, as we saw, the characteristic effects might be clinically relevant in double masked conditions but completely absent in routine practice where expectations are higher. The double masked trials of nicotine gum indicated that a double masked trial could have very low external validity: in double masked conditions there was a desirable effect of the nicotine in the gum, but in conditions similar to routine practice there was not.

Moreover, this provides a further reason to prefer 'active' controlled trials over 'placebo' controlled trials. If there are interactions, then the closer the strength of a study's non-characteristic features are to the strength of the non-characteristic features in routine practice, the more likely, intuitively speaking, the trial is to be externally valid. In an 'active' controlled trial, it is reasonable to assume that the strength of expectations is higher than in a placebo controlled trial. In an 'active' controlled trial, participants are given the choice between an established treatment and one that has shown initial promise (for example in phase I and II studies). This arguably makes the strength of the participants' expectations that they are taking an effective treatment in an 'active' controlled trial higher than they would be in a 'placebo' controlled trial and therefore closer to their level in the course of normal treatment.

To be sure, the expectations of the different groups in an 'active' controlled trial may not be the same. In fact empirical research (Chalmers 1997) suggests that in spite of the fact that new treatments are usually less effective than established treatments, people believe in the superior efficacy of the new intervention. Still, we would expect expectations regarding recovery in the two groups in an 'active' controlled trial to be higher than expectations in a placebo controlled trial and therefore closer to expectations 'in the wild'.

---

[122] This is not always the case. Sometimes the experimental treatment might have demonstrated effectiveness, and the demand for randomized trials be redundant (see Worrall, 2002). In other cases, treatments used in routine practice might not be based on a secure evidential foundation.

There is another way in which questioning the assumption of additivity could affect the external validity of trials. In routine practice, there are often, even usually, more than one 'treatment' prescribed. In addition to cases where patients take more than one drug, patients in routine clinical practice with hypertension, for example, could be prescribed an antihypertensive as well as given dietary and lifestyle (exercise) advice. In a well controlled RCT where the goal is to isolate the effect of certain characteristic features, the lifestyle and dietary advice will often be omitted or unnaturally standardized in order to isolate the effect of one or at most a few characteristic features. If, however, the changed diet and exercise regime interact with the drug somehow, then the external validity of the well-controlled trial will be compromised.

In short, if characteristic and non-characteristic features interact rather than add, then the goal of providing an absolute measure of effect size may inhibit the external validity of placebo controlled trials. I will now discuss the further sense in which placebo controlled trials do not provide a measure of absolute effect size.

8.4. Placebo Controls do not Provide a Standard Measure Against Which the Efficacy of The Characteristic Features can be Measured

There are reasons independent of the possibility of non-additivity to question the view that placebo controlled RCTs provide an 'absolute' measure of characteristic effectiveness. Unless the treatments used as 'placebo' controls are (a) legitimate, and (b) perform consistently, then placebo controlled trials cannot be said to provide an absolute measure of effectiveness.

I argued earlier (chapter 4) that many 'placebo' controls used in clinical trials are illegitimate, which is to say that they do not contain all and only the non-characteristic treatment features. Recall that the characteristic effects can roughly be thought of as calculated by subtracting the effects in the placebo group from the effects in the experimental group. But this 'equation' only holds if the non-characteristic features in the test treatment group and those in the control group are the same. There are certainly cases where 'placebo' controls are missing a non-characteristic feature of the experimental treatment. Quite obviously, if the 'placebo' control is missing a non-characteristic feature, then a measure of characteristic effectiveness will not be provided by a placebo controlled trial. More specifically, 'placebo' control treatments are often unsuccessful in deceiving the participants into thinking that they could be taking the experimental control treatment. Studies suggest that only a small fraction – less than

10% - of placebo controlled trials are successfully double masked (Fergusson et al. 2004; Hróbjartsson et al. 2007). The failure to keep a trial successfully double masked implies that the placebo control *lacks* one of the non-characteristic features that is part of the overall effect of the experimental treatment.

In other cases, the 'placebo' controls could have additional features that are not part of the experimental treatment. The example of olive oil used in placebo pill controls for cholesterol lowering drugs (before it was known that olive oil lowered cholesterol is one example of this (Golomb 1995). Subtracting the effects of an illegitimate 'placebo' control from the effects of the experimental treatment will not yield the effectiveness of (only) the characteristic features at all, let alone an *absolute* measure of the characteristic features.

Things get worse. Even legitimate placebos appear to vary widely in effectiveness (See chapter 6). In short, even where the placebo controls seem to contain all and only the non-characteristic features of the test treatment, and even where it appears as though these placebo control treatments are similar, the effects of these similar controls seems to vary. If so, then the measure of effect size would be relative to a particular trial and would, contrary to what is commonly held, not be an absolute measure of effect size.

The inconsistency of the placebo performance is not merely a distinction between the performance of the placebo control in a clinical trial compared with the performance of the placebo control in routine practice. If this were the case, then there would merely be a distinction between 'internal' absolute effect size and external absolute effect size (what you would see in routine treatment). The point would then turn into an admission that internal absolute effect size cannot be guaranteed to be (or even to approximate) external absolute effect size. The same placebo control can differ in its effects even within similar clinical trials.

In short, the claim that placebo controlled trials provide an absolute measure of effect size assumes that the placebo controls were legitimate. Since this assumption often cannot be made, it follows that placebo controlled trials cannot provide an absolute measure of characteristic effectiveness.

8.5. The Often Ignored Rival Hypotheses of Interactions and Illegitimate Placebos

'Scientific common sense' dictates that we must always demand that plausible rival hypotheses be ruled out before we accept the hypothesis at hand. In the case of the claim that placebo controlled trials provide a measure of the absolute effect of the characteristic features of the experimental treatment, three rival hypotheses must be ruled out. First, the possibility that the characteristic and non-characteristic features interact rather than add must be ruled out. Theoretical considerations (additivity is not the norm, especially in biochemistry), as well as several examples suggest that the rival hypothesis must be considered seriously rather than simply assumed to be false. Next, the hypothesis that the actual 'placebo' controls are legitimate must be evaluated. Given th previous discussion (chapter 4), it is fair to say that placebo controls are sometimes illegitimate and that therefore careful thought must be given before making claims about what a placebo controlled trial actually measures. Third, the possibility that the placebo control performed unusually well or poorly must also be ruled out. It is undoubtedly the case that these three rival hypotheses cannot often be ruled out. Therefore, placebo controlled trials cannot be said in general to provide a measure of absolute effect size.

Combined with the conclusion of the last chapter that placebo controlled trials are as assay *in*sensitive as 'active' controlled trials, this chapter makes the view that placebo controlled trials are superior to 'active' controlled trials difficult to uphold.

9. Chapter Nine. The Conceptual Foundation of Methodological Problems

> In the attack that has been launched on the merits of randomized controlled therapeutic trials, the target is sometimes cited in a confused or confusing manner. Someone who decries the idea of randomization may really object to the idea of double masking. Someone who complains about double masking may really be distressed about the types of agent, such as placebo, that are used as "controls". Attacks on placebo controls may really be directed at randomization, double-masking, concurrent comparison, or some other component of the evaluation process
>
> - *(Feinstein 1980)*

> What I do not believe – and this has been suggested –is that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect
>
> - *(Hill and Hill 1991)*

9.1. Common Sense Versus Mechanical Rules of Evidence

This thesis began with the claim that 'scientific common sense', the very general yet powerful claim that good evidence rules out plausible rival hypotheses, rather than any hard-and-fast rules of evidence or hierarchies, is (or should be) the bedrock of any account of scientific method. Mill's Methods, Bayesian Confirmation, and Popperian falsification, although outwardly very different, were shown to converge on this point.

A strict interpretation of the Evidence-Based Medicine (EBM) message, however, is that a 'hierarchy of evidence', with high-quality randomized trials at the top, is the best adjudicator of what counts as good evidence. No explicit reference is made to more fundamental principles that would guide the design and interpretation of the hierarchy. Yet, 'high quality' is taken to mean, among other things, that the trial was double masked, and, sometimes, placebo controlled. To be sure, many EBM proponents including the current director of Oxford's Centre for Evidence-Based Medicine (CEBM), Paul Glasziou, acknowledge that high-quality RCTs are not always better than other forms of evidence (Glasziou et al. 2007). Yet, at least officially and in many people's view, high-quality RCTs remain the best form of evidence.

This view, as I pointed out in the introduction to the thesis, leads to the paradox that many of the treatments that are surely most strongly supported by evidence have never been tested in randomized trials of any kind. These treatments include Automatic External Defibrullation to start a stopped heart, tracheostomy to open a blocked air

passage, the Heimlich manoeuvre to dislodge an obstruction in the breathing passage, rabies vaccines, and penicillin for the treatment of pneumonia.

Considerable previous literature in this area has focused on arguments that *randomization* controls for all baseline confounders, and that *randomized trials* in general are superior to other forms of evidence (Worrall 2002, 2007a, 2007b; Howson and Urbach 1993; Bluhm 2005; Lindley 1982, 1993; Penston 2003; Senn 1994, 2004). This thesis analyzes two additional features of randomized trials, namely

(1) double masking

(2) 'placebo' controlled trials and 'active' controlled trials.

However it is important to note that my arguments about these two potential features of clinical studies are quite independent of the arguments for and against randomization, as well as the arguments for and against RCTs. A trial could be double masked and placebo controlled but not randomized. Moreover, one might entirely reject the Bayesian arguments against randomization, and Worrall's arguments against RCTs while accepting my arguments that double masking is not always useful, and that 'active' controlled trials are methodologically equal to 'placebo' controlled trials.

Of course, my work impacts on the wider debate about evidential support in an indirect way. Double masking and placebo controls are potential features of RCTs but not observational studies. If these features always increase methodological quality, then there would be a further reason to prefer RCTs. If, on the other hand, they do not, then there are fewer potential reasons for preferring RCTs.

My first discovery was that the terms 'placebo control', 'double masking', and 'active controlled trial' are currently employed ambiguously. My first task was therefore to clarify the terms. Then, guided by 'scientific common sense' I found that double masking is not always of methodological value and that the arguments for the superiority of placebo controls are, on the whole, wanting. Throughout the course of my investigation I also found that double masking and the use of placebo controls hampered the external validity of clinical studies.

9.2. Questioning Double Masking as a Universal Virtue

The rationale for the view that double masking increases the quality of evidence is sound – it rules out expectations of participants and dispensing physicians as rival hypotheses for the outcome. Yet, it leads to the 'Phillip's Paradox', that the most effective treatments cannot be tested in double masked conditions. If a treatment's

characteristic features have characteristic effects, nobody will believe it is a placebo and attempts to keep the nature of the treatment 'masked' will fail. Moderately effective treatments, on the other hand, are more easily testable in double masked conditions.

Although double masking eliminates the potential confounding effects of participant expectation and investigator attitudes in theory, in practice these potential confounders may not be actual confounders, at least in certain cases. Moreover, the ethical requirement of informed consent makes double masking difficult, if not impossible, to achieve in practice. Outside of pharmaceutical drugs with mild effects, and that do not have strong side effects, double masking has proven, and is likely to continue to prove difficult if not impossible to implement successfully. Indeed most trials described as double masked do not remain successfully double masked. This means that being described as 'double masked' provides an illusion of increased methodological quality over 'open' trials. In short, open trials sometimes rule out as many plausible confounders as double masked trials, and they are easier to conduct.

Evaluating double masking by examining the circumstances where it did, in fact, rule out plausible rival hypotheses, or confounders, led to a resolution of the 'Phillip's Paradox'. The dramatic effects of certain treatments will, of course, swamp any potential effects of expectations and attitudes. In short, although *in principle* double masking rules out rival hypotheses, *in practice* the hypotheses it can rule out may not be plausible all the time.

9.3. The Problems with Placebo Versus Active Controls

A theme that immediately emerged when investigating the placebo is the ambiguity with which the term is used. My first task in evaluating the methodological role of placebo controls was to investigate definitions of placebos. Serious attempts to establish an acceptable conceptualization of placebos have been rare. Grünbaum's (1981/1986) account is a notable exception. I argue that Grünbaum's definition of placebos as treatments devoid of 'characteristic' features was a real advance over other definitions of placebos as 'inactive' or 'nonspecific' treatments. Not only are the other terms used to describe the placebo ambiguous (placebos can be both active and specific), but Grünbaum's insistence that what counts as a characteristic feature must be relativized to a therapeutic theory and target disorder avoids further confusion. The proverbial sugar pill, as Grünbaum noted, provides a *reductio* against the notion that there is such thing as a placebo *simpliciter* – sugar is not placebogenic for diabetes.

I have attempted to improve Grünbaum's account by adding a class of 'toxic agents'. Without this expansion, treatments with negative effects on the target disorder are classified as 'placebos', which is confusing. Second, I have defended Grünbaum's account against critiques by Waring (2003) and Greenwood (1997).

Waring argued that paradoxical drug responses (drugs that can have a positive effect on a target disorder for some people and a negative effect for the same disorder on others) forces Grünbaum into a contradiction: one and the same treatment can be both a toxic agent and a 'non-placebo'. I responded that it is clear that Grünbaum intended his definitional scheme to be relativized to patients. Hence, the same treatment can be a 'toxic agent' for some people, and a 'non-placebo' for others, and there is no contradiction. Greenwood argued that if positive effects can be explained pharmacologically (whether or not via any pharmacological factors deemed 'characteristic' by some therapeutic theory) then the treatment should *not* be classified as a placebo. Yet even pharmacological features, in some cases, can be 'placebogenic'. Both Greenwood and Waring claim that Grünbaum fails to emphasize that placebo effects must be psychological. This objection fails because psychological factors (i.e. in certain psychological treatments) should not be delivered an *a priori* classification as placebogenic.

In short, there is no such thing as 'the' placebo or 'the' placebo effect. More accurately, certain treatments are classified as 'placebos' by therapeutic theories for a particular ailment. Since different 'placebos' can be different (and have different effects), I argue that the term 'placebo' should be abandoned in favour of a precise definition of the control treatment in question.

Turning now to placebo controls, I claim that a 'legitimate' placebo control must rule out all rival hypotheses for apparent superiority of the experimental treatment *other than* its characteristic effects. I therefore defined 'legitimate placebo controls' as treatments that are no more and no less effective than the non-characteristic features of the test treatment. For example, a legitimate placebo control for treatment of depression involving Prozac would control for everything but the effects of fluoxetine. Quite obviously, the trial that employs an illegitimate placebo control will tend to lead to mistaken estimates of the characteristic effects of the experimental treatment. Although this definition seems obvious, careful consideration revealed that legitimate placebo controls will be difficult, even impossible to design outside the pharmaceutical setting. The non-characteristic features of many treatments with ill-defined (acupuncture) or

hopelessly intertwined (exercise) characteristic features will clearly be difficult to design, and assess for legitimacy. The actual 'placebo' controls used in trials testing these treatments may well be illegitimate, but without a better idea of what their characteristic features are, there is no way to tell. Hence the characteristic effects of these treatments are difficult to measure.

Even within the pharmaceutical setting, some actual 'placebo' controls used are illegitimate. For instance, a placebo control that fails to permit the study to remain successfully double masked, will not control for all non-characteristic features of the experimental treatment. In particular, such a trial will not control for participant expectations or dispenser attitudes. Illegitimate placebo controls lead to mistaken estimates of characteristic effectiveness of the experimental treatment, which can mislead patients, physicians, and policy makers.

The most important point to remember about the debate between placebo and 'active' controls is that 'placebo' controls, are treatments in and of their own right. They can, at least in principle, be as effective, or more effective than doing absolutely nothing. As such, they do not generally succeed at ruling out any special set of confounding factors or rival hypotheses. Rather, they suffer from the same problems, albeit perhaps to a lesser degree, as 'active' controls. More specifically, the 'assay sensitivity' arguments, which incidentally only apply to non-inferiority ACTs, fails precisely because they assume that placebo controls provide a reliable measure of effectiveness.

As for the argument that PCTs but not ACT provide a measure of absolute effect size, it based on the assumptions that placebo controls are legitimate, that they have constant effects, and that the non-characteristic and characteristic features of a treatment add rather than interact. These assumptions can rarely, if ever, be jointly held.

9.4. Weighing External Validity and Ethics Against Internal Validity

It is commonly claimed that placebo controls and double masking, if successfully implemented, increase the methodological quality of a trial. Although the bulk of my thesis calls into question the extent to which this claim is true, there is good reason to question the use of placebo controls and double masking even if they were as valuable as is sometimes claimed. The use of double masking and placebo controls makes clinical trials different from routine practice. This places the external validity of the trial

at risk. In particular, if characteristic and non-characteristic features do not add, then double masking and the element of doubt introduced by placebo controls could affect the generalizability of the trial results. The characteristic effectiveness in one trial could be very different from another, and perhaps more different from routine clinical practice. In short, choice to implement placebo controls and double masking might be taken without due weighing of potential benefits (which have been exaggerated) and risks (which have been understated).

Similar to the apparent tension between internal and external validity, there is an apparent tension between ethics and the methodology of placebo controls. Ethical considerations for the most part, mitigate against PCTs. Yet, in spite of these well recognized ethical constraints, PCTs are often advocated on methodological grounds. However, the supposed methodological advantages of PCTs over ACTs, namely that only the former are assay sensitive and provide a measure of absolute effect size, are, upon closer inspection mistaken to some degree. If this is so, then any tension between ethics and methodology dissolves, and the justification for PCTs where there is an established treatment becomes very weak indeed.

There are also two practical concerns that need to be considered. Pharmaceutical drugs with mild effects, are more easily supported by double blind, placebo controlled trials, regarding the double blind PCT as the gold standard will lead to these pharmaceutical drugs with mild effects being delivered *a priori* judgements of superior quality. Replacing hard and fast rules with 'scientific common sense', of course, provides the possibility of redressing this apparent inequity.

Next, what the average patient, practitioner, and policy maker wishes to know is which treatment, from among all available alternatives, is most effective (or cheapest or has the fewest side-effects, etc.) Placebo controlled trials do not provide this information as readily as ACTs. Hence, ACTs provide evidence that is of greater practical use than PCTs. Our standards of evidence should, in some way, reflect the aims we wish to achieve.

References

Ackerman, T. F. (2002), "Therapeutic beneficence and placebo controls", *Am J Bioeth* 2 (2):21-22.

Agarwal, A., R. Ranjan, S. Dhiraaj, A. Lakra, M. Kumar, and U. Singh (2005), "Acupressure for prevention of pre-operative anxiety: a prospective, randomised, placebo controlled study", *Anaesthesia* 60 (10):978-981.

Amanzio, M., A. Pollo, G. Maggi, and F. Benedetti (2001), "Response variability to analgesics: a role for non-specific activation of endogenous opioids", *Pain* 90 (3):205-215.

Anderson, James A. (2006), "The Ethics and Science of Placebo-Controlled Trials: Assay Sensitivity and the Duhem-Quine Thesis", *Journal of Medicine and Philosophy* 31:65-81.

Armitage, Peter, G. Berry, and J. N. S. Matthews (2002), *Statistical methods in medical research*. 4th ed. / P. Armitage, G. Berry, J.N.S. Matthews. ed. Oxford: Blackwell Science.

Aronson, J. K. (2007), "Concentration-effect and dose-response relations in clinical pharmacology", *Br J Clin Pharmacol* 63 (3):255-257.

Ashcroft, R. E. (2004), "Current epistemological problems in evidence based medicine", *Journal of Medical Ethics* 30:131035.

Baigent, C., R. Collins, P. Appleby, S. Parish, P. Sleight, and R. Peto (1998), "ISIS-2: 10 year survival among patients with suspected acute myocardial infarction in randomised comparison of intravenous streptokinase, oral aspirin, both, or neither. The ISIS-2 (Second International Study of Infarct Survival) Collaborative Group", *Bmj* 316 (7141):1337-1343.

Barton, S. (2000), "Which clinical studies provide the best evidence? The best RCT still trumps the best observational study", *Bmj* 321 (7256):255-256.

BBC News (1999), "Pill coating 'kills HIV'", in.

Beecher, H. K. (1955), "The powerful placebo", *J Am Med Assoc* 159 (17):1602-1606.

——— (1961), "Surgery as placebo. A quantitative study of bias", *Jama* 176:1102-1107.

——— (1962), "The placebo effect and sound planning in surgery", *Surg Gynecol Obstet* 114:507-509.

Begg, C., M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup (1996), "Improving the quality of reporting of randomized controlled trials. The CONSORT statement", *Jama* 276 (8):637-639.

Benedetti, F., and M. Amanzio (1997), "The neurobiology of placebo analgesia: from endogenous opioids to cholecystokinin", *Prog Neurobiol* 52 (2):109-125.

Benedetti, F., A. Pollo, L. Lopiano, M. Lanotte, S. Vighetti, and I. Rainero (2003), "Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses", *J Neurosci* 23 (10):4315-4323.

Benedetti, F., S. Vighetti, M. Amanzio, C. Casadio, A. Oliaro, B. Bergamasco, and G. Maggi (1998), "Dose-response relationship of opioids in nociceptive and neuropathic postoperative pain", *Pain* 74 (2-3):205-211.

Benedetti, Fabrizio, Luana Colloca, Leonardo Lopiano, and Michele Lanotte (2004), "Overt versus covert treatment for pain, anxiety, and Parkinson's disease", *The Lancet Neurology* 3 (November 2004).

Benedetti, Fabrizio, Innocenzo Rainero, and Antonella Pollo (2003), "New insights into placebo analgesia", *Current Opinion in Anaesthesiology* 16:515-519.
Benson, H., and D. P. McCallie, Jr. (1979), "Angina pectoris and the placebo effect", *N Engl J Med* 300 (25):1424-1429.
Benson, K., and A. J. Hartz (2000), "A comparison of observational studies and randomized, controlled trials", *N Engl J Med* 342 (25):1878-1886.
Bergmann, J. F., O. Chassany, and J. Gandiol (1994), "A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain", *Clinical Trials and Meta-Analysis* 29:41-47.
Birch, S., and R. N. Jamison (1998), "Controlled trial of Japanese acupuncture for chronic myofascial neck pain: assessment of specific and nonspecific effects of treatment", *Clin J Pain* 14 (3):248-255.
Black, N. (1996), "Why we need observational studies to evaluate the effectiveness of health care", *Bmj* 312 (7040):1215-1218.
Blackwelder, William C. (1982), ""Proving the Null Hypothesis" in Clinical Trials", *Controlled Clinical Trials* 3:345-353.
Bland, Martin (2000), *An introduction to medical statistics*. 3rd ed. ed. Oxford: Oxford University Press.
Block, A. E. (2007), "Costs and benefits of direct-to-consumer advertising: the case of depression", *Pharmacoeconomics* 25 (6):511-521.
Bluhm, R. (2005), "From hierarchy to network: a richer view of evidence for evidence-based medicine", *Perspect Biol Med* 48 (4):535-547.
Borgerson, K. (2005), "Evidence-based alternative medicine?" *Perspect Biol Med* 48 (4):502-515.
Branthwaite, A., and P. Cooper (1981), "Analgesic effects of branding in treatment of headaches", *Br Med J (Clin Res Ed)* 282 (6276):1576-1578.
Brighton, B., M. Bhandari, P. Tornetta, 3rd, and D. T. Felson (2003), "Hierarchy of evidence: from case reports to randomized controlled trials", *Clin Orthop Relat Res* (413):19-24.
Brown, B. S., T. Payne, C. Kim, G. Moore, P. Krebs, and W. Martin (1979), "Chronic response of rat brain norepinephrine and serotonin levels to endurance training", *J Appl Physiol* 46 (1):19-23.
Butler, Rob, Stuart Carney, Andrea Cipriani, John Geddes, John Hatcher, Jonathan Price, and Michael Von Korff (2003), "Depressive Disorders", in: Clinical Evidence.
Bylund, D. B., and A. L. Reed (2007), "Childhood and adolescent depression: why do children and adults respond differently to antidepressant drugs?" *Neurochem Int* 51 (5):246-253.
Canadian Task Force on the Periodic Health Examination (1979), "The periodic health examination", *Canadian Medical Association Journal* 121:1193-1254.
Cartwright, Nancy (1989), *Nature's capacities and their measurement*. Oxford: Clarendon.
——— (2007), "Are RCTs the Gold Standard?" in: Centre for the Philosophy of the Natural and Social Sciences.
——— (2007), *Hunting causes and using them : approaches in philosophy and economics*. Cambridge: Cambridge University Press.
Chalmers, I. (1997), "What is the prior probability of a proposed new treatment being superior to established treatments?" *Bmj* 314 (7073):74-75.

Choate, J. K., K. Kato, and R. M. Mohan (2000), "Exercise training enhances relaxation of the isolated guinea-pig saphenous artery in response to acetylcholine", *Exp Physiol* 85 (1):103-108.
Choudhuri, S., and L. G. Valerio, Jr. (2005), "Usefulness of studies on the molecular mechanism of action of herbals/botanicals: The case of St. John's wort", *J Biochem Mol Toxicol* 19 (1):1-11.
Clarke, Mike (2004), "Systematic Reviews and the Cochrane Collaboration", in, *The Cochrane Collaboration.*
Coats, T. L., D. G. Borenstein, N. K. Nangia, and M. T. Brown (2004), "Effects of valdecoxib in the treatment of chronic low back pain: results of a randomized, placebo-controlled trial", *Clin Ther* 26 (8):1249-1260.
Cochrane, A.L. (1972), "Effectiveness and efficiency: random reflections on health services", in, London: Nuffield Hospitals Trust.
Conan Doyle, Sir Arthur (1890), *The Sign of Four*. Edited by Project Gutenberg: Project Gutenberg.
Concato, J. (2004), "Observational versus experimental studies: what's the evidence for a hierarchy?" *NeuroRx* 1 (3):341-347.
Concato, J., N. Shah, and R. I. Horwitz (2000), "Randomized, controlled trials, observational studies, and the hierarchy of research designs", *N Engl J Med* 342 (25):1887-1892.
Connolly, S. J., R. Sheldon, K. E. Thorpe, R. S. Roberts, K. A. Ellenbogen, B. L. Wilkoff, C. Morillo, and M. Gent (2003), "Pacemaker therapy for prevention of syncope in patients with recurrent severe vasovagal syncope: Second Vasovagal Pacemaker Study (VPS II): a randomized trial", *Jama* 289 (17):2224-2229.
De Angelis, C., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. Van Der Weyden (2004), "Clinical trial registration: a statement from the International Committee of Medical Journal Editors", *N Engl J Med* 351 (12):1250-1251.
de Craen, A. J., P. J. Roos, A. Leonard de Vries, and J. Kleijnen (1996), "Effect of colour of drugs: systematic review of perceived effect of drugs and of their effectiveness", *Bmj* 313 (7072):1624-1626.
de Craen, A. J., J. G. Tijssen, J. de Gans, and J. Kleijnen (2000), "Placebo effect in the acute treatment of migraine: subcutaneous placebos are better than oral placebos", *J Neurol* 247 (3):183-188.
de Craen, A. J., J. G. Tijssen, and J. Kleijnen (1997), "Is there a need to control the placebo in placebo controlled trials?" *Heart* 77 (2):95-96.
Deupree, J. D., A. L. Reed, and D. B. Bylund (2007), "Differential effects of the tricyclic antidepressant desipramine on the density of adrenergic receptors in juvenile and adult rats", *J Pharmacol Exp Ther* 321 (2):770-776.
Devereaux, P. J., B. J. Manns, W. A. Ghali, H. Quan, C. Lacchetti, V. M. Montori, M. Bhandari, and G. H. Guyatt (2001), "Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials", *Jama* 285 (15):2000-2003.
Diks, J., D. Nio, V. Jongkind, M. A. Cuesta, J. A. Rauwerda, and W. Wisselink (2007), "Robot-assisted laparoscopic surgery of the infrarenal aorta : The early learning curve", *Surg Endosc.*
Dretske, Fred I. (1970), "Epistemic operators", *Journal of Philosophy* 67:1007-1023.
——— (1981), "The pragmatic dimension of knowledge", *Philosophical Studies* 40:363-378.

Dunn, A. L., M. H. Trivedi, J. B. Kampert, C. G. Clark, and H. O. Chambliss (2002), "The DOSE study: a clinical trial to examine efficacy and dose response of exercise as treatment for depression", *Control Clin Trials* 23 (5):584-603.
——— (2005), "Exercise treatment for depression: efficacy and dose response", *Am J Prev Med* 28 (1):1-8.
Dunnett, C. W., and M. Gent (1977), "Significance testing to establish equivalence between treatments, with special reference to data in the form of 2X2 tables", *Biometrics* 33 (4):593-602.
Edward, S. J., A. J. Stevens, D. A. Braunholtz, R. J. Lilford, and T. Swift (2005), "The ethics of placebo-controlled trials: a comparison of inert and active placebo controls", *World J Surg* 29 (5):610-614.
Ellenberg, S. S., and R. Temple (2000), "Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases", *Ann Intern Med* 133 (6):464-470.
Emanuel, E. J., and F. G. Miller (2001), "The ethics of placebo-controlled trials--a middle ground", *N Engl J Med* 345 (12):915-919.
Enkin, Murray W., Sholom Glouberman, Philip Groff, Alejandro Jadad, and Anita Stern (2005), "Beyond evidence: the complexity of maternity care", in, Toronto: The Clinamen Collaboration, 14.
Evans, Dylan (2003), *Placebo: The Belief Effect*: Harper Collins.
Fabre, L. F., and H. P. Putman, 3rd (1987), "A fixed-dose clinical trial of fluoxetine in outpatients with major depression", *J Clin Psychiatry* 48 (10):406-408.
FDA (1996), "F.D.A Consumer", 30 (5).
——— (1998), "Guidance for Industry", in U.S. Department of Health and Human Services (ed.).
——— (2005), "Part 314: Applications for FDA Approval to Market a New Drug", in: United States Food and Drug Administration.
Feinstein, A. R. (1980), "Should placebo-controlled trials be abolished?" *Eur J Clin Pharmacol* 17 (1):1-4.
——— (2002), "Post-therapeutic response and therapeutic "style": re-formulating the "placebo effect"", *J Clin Epidemiol* 55 (5):427-429.
Fergusson, D., K. C. Glass, D. Waring, and S. Shapiro (2004), "Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials", *Bmj* 328 (7437):432.
Fisher, Ronald A. (1947), *The design of experiments*. 4th ed. Edinburgh: Oliver & Boyd.
Freed, C. R., P. E. Greene, R. E. Breeze, W. Y. Tsai, W. DuMouchel, R. Kao, S. Dillon, H. Winfield, S. Culver, J. Q. Trojanowski, D. Eidelberg, and S. Fahn (2001), "Transplantation of embryonic dopamine neurons for severe Parkinson's disease", *N Engl J Med* 344 (10):710-719.
Freedman, B., K. C. Glass, and C. Weijer (1996), "Placebo orthodoxy in clinical research. II: Ethical, legal, and regulatory myths", *J Law Med Ethics* 24 (3):252-259.
Freedman, B., C. Weijer, and K. C. Glass (1996), "Placebo orthodoxy in clinical research. I: Empirical and methodological myths", *J Law Med Ethics* 24 (3):243-251.
Friedman, J. H. (2004), "Randomized, double-blind, placebo-controlled trials: the gold standard?" *Med Health R I* 87 (9):262-263.
Gardner, P. (2003), "Distorted Packaging: Marketing Depression as Illness, Drugs as Cure", *Journal of Medical Humanities* 24 (1-2):105-130.

Gillies, Donald (1986), "In defense of the Popper-Miller argument", *Philosophy of Science* 53 (111).
——— (1990), "The Turing-Good Weight of Evidence Function and Popper's Measure of the Severity of a Test", *British Journal for the Philosophy of Science* 41:143-146.
——— (1998), "Confirmation Theory", in D. M. Gabbay and Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, 135-167.
——— (2005), "An Action-Related Theory of Causality", *British Journal for the Philosophy of Science*.
Glasziou, P., I. Chalmers, M. Rawlins, and P. McCulloch (2007), "When are randomised trials unnecessary? Picking signal from noise", *Bmj* 334 (7589):349-351.
Golomb, B. A. (1995), "Paradox of placebo effect", *Nature* 375 (6532):530.
Gomberg-Maitland, M., L. Frison, and J. L. Halperin (2003), "Active-control clinical trials to establish equivalence or noninferiority: methodological and statistical concepts linked to quality", *Am Heart J* 146 (3):398-403.
Gøtzsche, P. C. (1994), "Is there logic in the placebo?" *Lancet* 344 (8927):925-926.
Gragoudas, E. S., A. P. Adamis, E. T. Cunningham, Jr., M. Feinsod, and D. R. Guyer (2004), "Pegaptanib for neovascular age-related macular degeneration", *N Engl J Med* 351 (27):2805-2816.
Green, S, and J Higgins *Glossary. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5*, May 2005 2005 [cited 23 November 2006. Available from www.cochrane.org/resources/glossary.htm, accessed 1 December 2007.
Greene, W. L., J. Concato, and A. R. Feinstein (2000), "Claims of equivalence in medical research: are they supported by the evidence?" *Ann Intern Med* 132 (9):715-722.
Greenhalgh, Trisha (2006), *How to read a paper : the basics of evidence-based medicine*. 3rd ed. ed. Malden, Mass.: BMJ Books/Blackwell Pub.
Greenwood, John D. (1996), "Freud's "Tally" Argument, Placebo Control Treatments, and the Evaluation of Psychotherapy", *Philosophy of Science* 62:605-621.
——— (1997), "Placebo Control Treatments and the Evaluation of Psychotherapy: A Reply to Grunbaum and Erwin", *Philosophy of Science* 64 (September):497-510.
Grossman, Jason, and Fiona MacKenzie (2005), "The Randomized Controlled Trial: gold standard, or merely standard?" *Perspectives in Biology and Medicine* 48 (4):516-534.
Grünbaum, A. (1981), "The placebo concept", *Behav Res Ther* 19 (2):157-167.
——— (1986), "The placebo concept in medicine and psychiatry", *Psychol Med* 16 (1):19-38.
Guyatt, G. (1991), "Evidence-based medicine", *Americal College of Physicians Journal Club* 114:A16.
Gyotoku, T., L. Aurelian, and A. R. Neurath (1999), "Cellulose acetate phthalate (CAP): an 'inactive' pharmaceutical excipient with antiviral activity in the mouse model of genital herpesvirus infection", *Antivir Chem Chemother* 10 (6):327-332.
Haake, M., H. H. Muller, C. Schade-Brittinger, H. D. Basler, H. Schafer, C. Maier, H. G. Endres, H. J. Trampisch, and A. Molsberger (2007), "German Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups", *Arch Intern Med* 167 (17):1892-1898.

Hall, S. D., Z. Wang, S. M. Huang, M. A. Hamman, N. Vasavada, A. Q. Adigun, J. K. Hilligoss, M. Miller, and J. C. Gorski (2003), "The interaction between St John's wort and an oral contraceptive", *Clin Pharmacol Ther* 74 (6):525-535.

Hamilton, M. (1967), "Development of a rating scale for primary depressive illness", *Br J Soc Clin Psychol* 6 (4):278-296.

Harbour, Robin T, ed. (2008), *SIGN 50: A guideline developer's handbook.* Edited by Scottish Intercollegiate Guidelines Network. Edinburgh: NHS Quality Improvement Scotland.

Hayes, M. A., A. C. Timmins, E. H. Yau, M. Palazzo, C. J. Hinds, and D. Watson (1994), "Elevation of systemic oxygen delivery in the treatment of critically ill patients", *N Engl J Med* 330 (24):1717-1722.

Haynes, Brian R (2002), "What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to?" *BMC Health Serv Res* 2:3.

Healy, D. (2006), "Did regulators fail over selective serotonin reuptake inhibitors?" *Bmj* 333 (7558):92-95.

Healy, David (2004), *Let Them Eat Prozac.* New York: New York University Press.

Heckerling, P. S. (2006), "Placebo surgery research: a blinding imperative", *J Clin Epidemiol* 59 (9):876-880.

Hempel, Carl Gustav (1966), *Philosophy of natural science, (Prentice-Hall Foundations of philosophy series.)*: Englewood Cliffs (N.J.): Prentice-Hall.

Higgins, Julia (2005), *RCTs offer the best evidence.* London.

Hill, Austin Bradford, and I. D. Hill (1991), *Bradford Hill's principles of medical statistics.* 12th ed. ed: Edward Arnold.

Holmes, David, Stuart Murray, Amelie Perron, and Genevieve Rail (2006), "Deconstructing the evidence-based discourse in health sciences: truth, power and fascism", *Int J Evid Based Healthc* 4:180-186.

Howson, Colin (2000), *Induction and the justification of belief : Hume's problem.* Oxford , New York: Clarendon Press, Oxford University Press.

Howson, Colin, and Peter Urbach (1993), *Scientific reasoning : the Bayesian approach.* 2nd ed. Illinois: Open Court.

Hróbjartsson, A. (1996), "The uncontrollable placebo effect", *Eur J Clin Pharmacol* 50 (5):345-348.

——— (2002), "What are the main methodological problems in the estimation of placebo effects?" *J Clin Epidemiol* 55 (5):430-435.

Hróbjartsson, A., E. Forfang, M. T. Haahr, B. Als-Nielsen, and S. Brorson (2007), "Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding", *Int J Epidemiol.*

Hróbjartsson, A., and P. C. Gøtzsche (2001), "Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment", *N Engl J Med* 344 (21):1594-1602.

——— (2004a), "Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment", *J Intern Med* 256 (2):91-100.

——— (2004b), "Placebo interventions for all clinical conditions", *Cochrane Database Syst Rev* (3):CD003974.

Hughes, J. R., S. B. Gulliver, G. Amori, G. C. Mireault, and J. F. Fenwick (1989), "Effect of instructions and nicotine on smoking cessation, withdrawal symptoms and self-administration of nicotine gum", *Psychopharmacology (Berl)* 99 (4):486-491.

Huskisson, E. C. (1974), "Simple analgesics for arthritis", *Br Med J* 4 (5938):196-200.
Hwang, Irving K., and Toshinhiko Morikawa (1999), "Design Issues in Noninferiority/Equivalence Trials", *Drug Information Journal* 33:1205-1218.
ICH (2000), "ICH (International Conference on Harmonization) Harmonized Tripartite Guideline. Choice of Control Group and Related Issues in Clinical Trials. E 10", in Department of Health and Human Services (ed.): Centre for Biologics Evaluation and Research.
Jadad, Alejandro (1998), *Randomized Controlled Trials*. London: BMJ Books.
Jefferson, Thomas, and Frank Irwin (1975), *Letters of Thomas Jefferson*. Tilton, N.H.: Sanbornton Bridge Press.
Kaptchuk, T. J. (1998), "Intentional Ignorance: A History of Blind Assessment and Placebo Controls in Medicine", *Bulletin of the History of Medicine* 72.3:389-433.
——— (2001), "The double-blind, randomized, placebo-controlled trial: gold standard or golden calf?" *J Clin Epidemiol* 54 (6):541-549.
Kaptchuk, T. J., W. B. Stason, R. B. Davis, A. R. Legedza, R. N. Schnyer, C. E. Kerr, D. A. Stone, B. H. Nam, I. Kirsch, and R. H. Goldman (2006), "Sham device v inert pill: randomised controlled trial of two placebo treatments", *Bmj* 332 (7538):391-397.
Katz, R. D., J. A. Taylor, G. D. Rosson, P. R. Brown, and N. K. Singh (2006), "Robotics in plastic and reconstructive surgery: use of a telemanipulator slave robot to perform microvascular anastomoses", *J Reconstr Microsurg* 22 (1):53-57.
Kienle, G. S., and H. Kiene (1997), "The powerful placebo effect: fact or fiction?" *J Clin Epidemiol* 50 (12):1311-1318.
Kirsch, I. (2000), "Are drug and placebo effects in depression additive?" *Biol Psychiatry* 47 (8):733-735.
——— (2002), "Antidepressants and Placebos: Secrets, Revelations, and Unanswered Questions", *Prevention & Treatment* 5.
——— (2002), "Yes, There Is a Placebo Effect, but Is Ther a Powerful Antidepressant Effect?" *Prevention and Treatment* 5 (22).
——— (2003), "Hidden Administration as Ethical Alternatives to the Balanced Placebo Design", *Prevention & Treatment* 6.
——— (2004), "Conditioning, expectancy, and the placebo effect: comment on Stewart-Williams and Podd (2004)", *Psychol Bull* 130 (2):341-343; discussion 344-345.
——— (2005), "Placebo psychotherapy: Synonym or oxymoron?" *J Clin Psychol.*
Kirsch, I., and Thomas Moore (2002), "The Emperor's New Drugs: An Analysis of Antidepressant Medication Data Submitted to the U.S. Food and Drug Administration", *Prevention & Treatment* 5.
Kirsch, I., and Guy Sapirstein (1998), "Listening to Prozac but Hearing Placebo: A Meta-Analysis of Antidepressant Medication", *Prevention & Treatment* I.
Kleijnen, J., A. J. de Craen, J. van Everdingen, and L. Krol (1994), "Placebo effect in double-blind clinical trials: a review of interactions with medications", *Lancet* 344 (8933):1347-1349.
Klein, Donald F. (1998), "Listening to Meta-Analysis but Hearing Bias", *Prevention & Treatment* 1 (Article 0006c).
Koch, G., U. Johansson, and E. Arvidsson (1980), "Radioenzymatic determination of epinephrine, norepinephrine, and dopamine in 0.1 mL plasma samples: plasma

catecholamine response to submaximal and near maximal exercise", *The Journal of Clinical Chemistry and Clinical Biochemistry*:367-372.
Lasagna, L. (1955), "The controlled clinical trial: theory and practice", *J Chronic Dis* 1 (4):353-367.
Lee, J., M. Dodd, S. Dibble, and D. Abrams (2008), "Review of Acupressure Studies for Chemotherapy-Induced Nausea and Vomiting Control", *J Pain Symptom Manage*.
Leibovici, L. (2001), "Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial", *Bmj* 323 (7327):1450-1451.
Leopold, S. S., W. J. Warme, E. Fritz Braunlich, and S. Shott (2003), "Association between funding source and study outcome in orthopaedic research", *Clin Orthop Relat Res* (415):293-301.
Lesaffre, E., E. Bluhmki, F. Wang-Clow, S. Berioli, T. Danays, N. L. Fox, and F. Van de Werf (2001), "The general concepts of an equivalence trial, applied to ASSENT-2, a large-scale mortality study comparing two fibrinolytic agents in acute myocardial infarction", *Eur Heart J* 22 (11):898-902.
Levine, J. D., and N. C. Gordon (1984), "Influence of the method of drug administration on analgesic response", *Nature* 312 (5996):755-756.
Lewis, David (1996), "Elusive Knowledge", *Australasian Journal of Philosophy* 74:549-567.
Lewis, J. A., B. Jonsson, G. Kreutz, C. Sampaio, and B. van Zwieten-Boot (2002), "Placebo-controlled trials and the Declaration of Helsinki", *Lancet* 359 (9314):1337-1340.
Lindley, D.V. (1982), "The Role of Randomization in Inference", *PSA* 2:431-446.
——— (1993), "The analysis of experimental data: the appreciation of tea and wine", *Teach Statistics* 15:22-25.
Lundeberg, T., and I. Lund (2007), "Are reviews based on sham acupuncture procedures in fibromyalgia syndrome (FMS) valid?" *Acupunct Med* 25 (3):100-106.
Mackie, J.L. (1974), *The Cement of the Universe*. Oxford: Clarendon Press.
Max, M. B. (1994), "Divergent traditions in analgesic clinical trials", *Clin Pharmacol Ther* 56 (3):237-241.
McCann, I. L., and D. S. Holmes (1984), "Influence of aerobic exercise on depression", *J Pers Soc Psychol* 46 (5):1142-1147.
Mill, John Stuart (1843[1973]), *A system of logic, ratiocinative and inductive : being a connected view of the principles of evidence and the methods of scientific investigation*. Edited by Robson, *Collected works of John Stuart Mill ; v.7-8*. Toronto: University of Toronto Press.
Miller, F. G., and H. Brody (2002), "What makes placebo-controlled trials unethical?" *Am J Bioeth* 2 (2):3-9.
Miller, J. N., G. A. Colditz, and F. Mosteller (1989), "How study design affects outcomes in comparisons of therapy. II: Surgical", *Stat Med* 8 (4):455-466.
Miller, M. F., and N. W. Chilton (1980), "The effect of an oxygenating agent upon recurrent aphthous stomatitis -- a double-blind clinical trial", *Pharmacol Ther Dent* 5 (3-4):55-58.
Mitchell, S. H., C. L. Laurent, and H. de Wit (1996), "Interaction of expectancy and the pharmacological effects of d-amphetamine: subjective effects and self-administration", *Psychopharmacology (Berl)* 125 (4):371-378.

Moerman, D. E. (1983), "General Medical Effectiveness and Human Biology: Placebo Effects in the Treatment of Ulcer Disease", *Medical Anthropology Quarterly* 14 (4):3+13-16.
——— (2000), "Cultural Variations in the Placebo Effect: Ulcers, Anxiety, and Blood Pressure", *Medical Anthropology Quarterly* 14 (1):51-72.
Moerman, D. E., and W. B. Jonas (2002), "Deconstructing the placebo effect and finding the meaning response", *Ann Intern Med* 136 (6):471-476.
Moher, D., B. Pham, A. Jones, D. J. Cook, A. R. Jadad, M. Moher, P. Tugwell, and T. P. Klassen (1998), "Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?" *Lancet* 352 (9128):609-613.
Moher, D., K. F. Schulz, and D. G. Altman (2001), "The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials", *Lancet* 357 (9263):1191-1194.
Moncrieff, J. (2003), "A comparison of antidepressant trials using active and inert placebos", *Int J Methods Psychiatr Res* 12 (3):117-127.
Moncrieff, J., and I. Kirsch (2005), "Efficacy of antidepressants in adults", *Bmj* 331 (7509):155-157.
Moncrieff, J., and S. Wessely (1998), "Active placebos in antidepressant trials", *Br J Psychiatry* 173:88.
Moncrieff, J., S. Wessely, and R. Hardy (2004), "Active placebos versus antidepressants for depression", *Cochrane Database Syst Rev* (1):CD003012.
Montgomery, G. H., and I. Kirsch (1997), "Classical conditioning and the placebo effect", *Pain* 72 (1-2):107-113.
Montori, V. M., M. Bhandari, P. J. Devereaux, B. J. Manns, W. A. Ghali, and G. H. Guyatt (2002), "In the dark: the reporting of blinding status in randomized controlled trials", *J Clin Epidemiol* 55 (8):787-790.
Morris, David B. (1997), "Placebo, Pain, and Belief: A Biocultural Model", in Anne Harrington (ed.), *The Placebo Effect: an interdisciplinary exploration*, Cambridge, MA: Harvard University Press, 187-207.
Moseley, J. B., K. O'Malley, N. J. Petersen, T. J. Menke, B. A. Brody, D. H. Kuykendall, J. C. Hollingsworth, C. M. Ashton, and N. P. Wray (2002), "A controlled trial of arthroscopic surgery for osteoarthritis of the knee", *N Engl J Med* 347 (2):81-88.
Motamed, C., X. Mazoit, K. Ghanouchi, F. Guirimand, K. Abhay, T. Lieutaud, S. Bensaid, C. Fernandez, and P. Duvaldestin (2000), "Preemptive intravenous morphine-6-glucuronide is ineffective for postoperative pain relief", *Anesthesiology* 92 (2):355-360.
N.I.H (2004), " Directives for Human Experimentation: NUREMBERG CODE", in National Institute of Health (ed.).
——— (2006), *ClinicalTrials.gov*. N.I.H, July 2006 2006 [cited 23 November 2006 2006]. Available from http://www.clinicaltrials.gov.
Ney, P. G., C. Collins, and C. Spensor (1986), "Double blind: double talk or are there ways to do better research", *Med Hypotheses* 21 (2):119-126.
NIH (1997), "Acupuncture. NIH Consensus Statement", in NIH Consensus Department (ed.), London: National Institutes of Health, 1-34.
OED (2007), *The Oxford English Dictionary* (2nd) [Online]. Oxford University Press 1989 [cited 16 August 2007 2007]. Available from www.oed.com.
Olanow, C. W., C. G. Goetz, J. H. Kordower, A. J. Stoessl, V. Sossi, M. F. Brin, K. M. Shannon, G. M. Nauert, D. P. Perl, J. Godbold, and T. B. Freeman (2003), "A

double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson's disease", *Ann Neurol* 54 (3):403-414.
Onwude, J. (2005), "Stress incontinence", *Clin Evid* (14):2365-2382.
Papineau, D. (1994), "The virtues of randomization", *British Journal for the Philosophy of Science* 45:437-450.
Paterson, C., and P. Dieppe (2005), "Characteristic and incidental (placebo) effects in complex interventions such as acupuncture", *Bmj* 330 (7501):1202-1205.
Pearl, Judea (2000), *Causality : models, reasoning, and inference*. Cambridge: Cambridge University Press.
Penston, James (2003), *Fact and Fiction in Medical Research: The Large-Scale Randomised Trial*. London: The London Press.
Peto, R., R. Collins, and R. Gray (1995), "Large-scale randomized evidence: large, simple trials and overviews of trials", *J Clin Epidemiol* 48 (1):23-40.
Pfrunder, A., M. Schiesser, S. Gerber, M. Haschke, J. Bitzer, and J. Drewe (2003), "Interaction of St John's wort with low-dose oral contraceptive therapy: a randomized controlled trial", *Br J Clin Pharmacol* 56 (6):683-690.
Phillips, Bob, Chris Ball, Dave Sackett, Dough Badenoch, Sharon Straus, Brian Haynes, and Martin Dawes (2001), "Oxford Centre for Evidence-based Medicine Levels of Evidence ", in CEBM (ed.): CEBM.
Piaggio, G., D. R. Elbourne, D. G. Altman, S. J. Pocock, and S. J. Evans (2006), "Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement", *Jama* 295 (10):1152-1160.
Pocock, S. J., and D. R. Elbourne (2000), "Randomized trials or observational tribulations?" *N Engl J Med* 342 (25):1907-1909.
Popper, Karl R. (1968), *The logic of scientific discovery*. Rev. ed, *Radius Books*. London: Hutchinson. Original edition, 1959.
——— (1969), *Conjectures and refutations : the growth of scientific knowledge*. London: Routledge & K. Paul.
Rickels, K., W. T. Smith, V. Glaudin, J. B. Amsterdam, C. Weise, and G. P. Settle (1985), "Comparison of two dosage regimens of fluoxetine in major depression", *J Clin Psychiatry* 46 (3 Pt 2):38-41.
Roethlisberger, F. J., and W. J. Dickson (1939), *Management and the Worker*: Harvard University Press.
Rohsenow, D. J., and G. A. Marlatt (1981), "The balanced placebo design: methodological considerations", *Addict Behav* 6 (2):107-122.
Rosenthal, Robert, and Lenore F. Jacobson (1992), *Pygmalion in the classroom : teacher expectation and pupils' intellectual development*. New York: Irvington Publishers.
Ross, D. F., and R. O. Pihl (1989), "Modification of the balanced-placebo design for use at high blood alcohol levels", *Addict Behav* 14 (1):91-97.
Rossouw, J. E., G. L. Anderson, R. L. Prentice, A. Z. LaCroix, C. Kooperberg, M. L. Stefanick, R. D. Jackson, S. A. Beresford, B. V. Howard, K. C. Johnson, J. M. Kotchen, and J. Ockene (2002), "Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial", *Jama* 288 (3):321-333.
Sackett, D. L. (1991), *Clinical epidemiology : a basic science for clinical medicine*. 2nd ed. / David L. Sackett...[et al.] ed: Little, Brown.
——— (2004), "Turning a blind eye: why we don't test for blindness at the end of our trials", *Bmj* 328 (7448):1136.
Savage, J (1976), "On Rereading R. A. Fisher", *The Annals of Statistics* 4:441-500.

Schlesinger, George N. (1995), "Measuring Degrees of Confirmation", *Analysis* 55.3:208-212.
Schulz, K. F., I. Chalmers, and D. G. Altman (2002), "The landscape and lexicon of blinding in randomized trials", *Ann Intern Med* 136 (3):254-259.
Schulz, K. F., I. Chalmers, R. J. Hayes, and D. G. Altman (1995), "Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials", *Jama* 273 (5):408-412.
Senn, S. J. (1994), "Fisher's game with the devil", *Stat Med* 13 (3):217-230.
——— (2004), "Controversies concerning randomization and additivity in clinical trials", *Stat Med* 23 (24):3729-3753.
——— (2005), "Active Control Equivalence Studies", in B. S. Everitt and C. R. Palmer (eds.), *Encyclopaedic Companion to Medical Statistics*: Hodder Arnold, 19-22.
——— (2006), "Giving Chance its due: the dangers of over-interpreting differences in observed placebo response", in: Department of Statistics, University of Glasgow, 12.
——— (2007a), *Statistical Issues in Drug Development*. 2nd ed: Wiley: Hoboken.
Shaffer, Jonathan (2001), "Knowledge, relevant alternatives and missed clues", *Analysis* 61 (3):202-208.
Shapiro, A., and Louis A. Morris (1978), "The Placebo Effect in Medical and Psychological Therapies", in Sol L. Garfield and Allen E. Bergin (eds.), *Handbook of Psychotherapy and Behavioural Change: An Empirical Analysis*, New York: John Wiley & Sons, 369-410.
Shapiro, A., and E. Shapiro (1997), "The Placebo. Is it Much Ado About Nothing?" in Anne Harrington (ed.), *The Placebo Effect: An Interdisciplinary Exploration.*
Shapiro, Stan (2004), "Widening the field of vision", *BMJ* Rapid Responses.
Smith, G. C., and J. P. Pell (2003), "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials", *Bmj* 327 (7429):1459-1461.
Sprott, H., R. E. Gay, B. A. Michel, and S. Gay (2006), "Influence of ibuprofen-arginine on serum levels of nitric oxide metabolites in patients with chronic low back pain--a single-blind, placebo controlled pilot trial (ISRCTN18723747)", *J Rheumatol* 33 (12):2515-2518.
Straus, Sharon E., W. Scott Richardson, and R. Brian Haynes (2005), *Evidence-Based Medicine: How to Practice and Teach EBM.* 3rd ed. London: Elsevier: Churchill Livingstone.
Streitberger, K., and J. Kleinhenz (1998), "Introducing a placebo needle into acupuncture research", *Lancet* 352 (9125):364-365.
Sullivan, P. F., K. S. Kendler, and M. C. Neale (2003), "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies", *Arch Gen Psychiatry* 60 (12):1187-1192.
Suzuki, N., A. Hattori, S. Suzuki, and Y. Otake (2007), "Development of a surgical robot system for endovascular surgery with augmented reality function", *Stud Health Technol Inform* 125:460-463.
Takala, J., E. Ruokonen, N. R. Webster, M. S. Nielsen, D. F. Zandstra, G. Vundelinckx, and C. J. Hinds (1999), "Increased mortality associated with growth hormone treatment in critically ill adults", *N Engl J Med* 341 (11):785-792.
Temple, Robert, and Susan Ellenberg (2000), "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments: Part 1: Ethical and Scientific Issues", *Annals of Internal Medicine* 133 (6):455-463.

ter Riet, G., A. J. de Craen, A. de Boer, and A. G. Kessels (1998), "Is placebo analgesia mediated by endogenous opioids? A systematic review", *Pain* 76 (3):273-275.
Torgerson, D. J., and B. Sibbald (1998), "Understanding controlled trials. What is a patient preference trial?" *Bmj* 316 (7128):360.
Tramer, M. R., D. J. Reynolds, R. A. Moore, and H. J. McQuay (1998), "When placebo controlled trials are essential and equivalence trials are inadequate", *Bmj* 317 (7162):875-880.
Travers, J., S. Marsh, M. Williams, M. Weatherall, B. Caldwell, P. Shirtcliffe, S. Aldington, and R. Beasley (2007), "External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?" *Thorax* 62 (3):219-223.
Urbach, P (1985), "Randomization and the Design of Experiments", *Philosophy of Science* 52:256-273.
US Preventive Services Task Force (1996), "Guide to Clinical Preventive Services", in US Department of Health and Human Services (ed.): DHHS.
Van Spall, H. G., A. Toren, A. Kiss, and R. A. Fowler (2007), "Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review", *Jama* 297 (11):1233-1240.
Walach, Harald, Torkel Falkenberg, Vinjar Fonnebo, George Lewith, and Wayne B. Jonas (2006), "Circular instead of hierarchical: methodological principles for the evaluation of complex interventions", *BMC Research Methodology* 6 (29).
Waring, D. (2003), "Paradoxical drug response and the placebo effect: a discussion of Grunbaum's definitional scheme", *Theor Med Bioeth* 24 (1):5-17.
Weijer, C., and K. C. Glass (2002), "The ethics of placebo-controlled trials", *N Engl J Med* 346 (5):382-383.
Weijer, C., and P. B. Miller (2004), "When are research risks reasonable in relation to anticipated benefits?" *Nat Med* 10 (6):570-573.
WMA (1964), "Declaration of Helsinki", in, Helsinki: World Medical Association.
——— (2001), "Note of clarification on paragraph 29 of the WMA Declaration of Helsinki, Geneva", in World Medical Association (ed.): World Medical Association.
Woolery, A., H. Myers, B. Sternlieb, and L. Zeltzer (2004), "A yoga intervention for young adults with elevated symptoms of depression", *Altern Ther Health Med* 10 (2):60-63.
Worn, H. (2006), "Computer- and robot-aided head surgery", *Acta Neurochir Suppl* 98:51-61.
Worrall, John (2002), "*What* Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69 (Supplement):S316-S330.
——— (2007a), "Why There's no Cause to Randomize", *British Journal for the Philosophy of Science* 58 (3):451-488.
——— (2007b), "Evidence in Medicine", *Compass* forthcoming.
Yaphe, J., R. Edman, B. Knishkowy, and J. Herman (2001), "The association between funding by commercial interests and study outcome in randomized controlled drug trials", *Fam Pract* 18 (6):565-568.
Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight (1985), "Beta blockade during and after myocardial infarction: an overview of the randomized trials", *Prog Cardiovasc Dis* 27 (5):335-371.